Identifying the Most Important Factors for Predicting Threat Detection Performance in Aviation Security X-Ray Screening

Thesis
presented to the Faculty of Arts
of
the University of Zurich
for the degree of Doctor of Philosophy

by
Anton Bolfing

of Schwyz SZ

Accepted in the fall semester 2008 on the recommendation of
Prof. Dr. Wolfgang Marx and Prof. Dr. Martin Kleinmann

ETH Press, Zurich
2008

Thesis advisors

Prof. Dr. Wolfgang Marx

Prof. Dr. Martin Kleinmann

Author

Anton Bolfing

Identifying the Most Important Factors for Predicting Threat Detection Performance in Aviation Security X-Ray Screening

Abstract

Together with the tremendous growth of civil aviation in the last decades, the importance of aviation security and its public perception have dramatically increased. Despite of large technological progress in X-ray imaging such as high-resolution image quality and image enhancement features, the final decision whether a luggage piece will enter an aircraft or not is always made by a human operator. Therefore, human factors are still the essential key element in aviation security worldwide. Even if no technological equipment will replace human operators from the X-ray screening task in the near future, EU audits and covert tests have sometimes shown serious operational deficiencies at the checkpoints.

The core subject of this thesis is mainly about image based factors that affect threat detection performance of human screeners that operate the X-ray equipments at security checkpoints. Schwaninger, Hardmeier, and Hofer (2004) have identified three image based factors: View Difficulty, Superposition and Bag Complexity. In my thesis I developed computational image measurements to automatically estimate such image based factors. Therefore a series of adaptations to these three concepts have been necessary for computational implementation. In this thesis I finally ended up with the following new concepts: Threat Category, View Difficulty, Superposition, Opacity and Clutter constituting Bag Complexity, and Bag Size as a new factor. Applying statistical models to these image based factors allows investigating the concerted impact of these image based factors on threat detection. All reported studies revealed that it is possible to predict average detection performance on

a single image quite well solely based on computationally accessible image properties.

Since detection performance in visual search tasks depends on the stimulus material (image based factors) but also on human factors, human factors should not be neglected in a comprehensive model. Therefore in later works also Training and Age were included into our statistical models. This allows comparing the impact of image based factors and human factors. Applications of a computational model for threat image projection systems and for adaptive computer-based training are discussed.

Contents

Personal Contributions

Large parts of the work at hand has been previously published in peer-reviewed journals and conference papers. In some of them I am listed as first author and in others not. In any case I was the one who wrote them. Some of them have been revised by my mentor Adrian Schwaninger, some haven't. Further, for the most part, the conducted analyses were accomplished by myself. Generally spoken it is my work, entirely. The reasons for me not to be the first author in every work is, that Adrian Schwaninger was the person who was responsible for project acquisitions, without which probably none of the presented studies ever could have been initiated. In those cases where I did not accomplish the bigger parts of the statistical analyses, a small part, the reason is our VICOREG quality management policy. Because of the sensitivity and the impact of our studies, we have double check all statistical analyses we conduct. In those cases where I was not the one doing the primary analysis I was in charge of the double checking. At this occasion I would like to anticipate my acknowledgment to Tobias Halbherr for his great job doing the primary data analysis on very large data bases, a real challenge.

Acknowledgments

Completing this doctoral work has been a wonderful and often overwhelming experience. It is hard to know whether it has been grappling with the cognitive psychology itself which has been the real learning experience or getting into the subject of aviation security, or grappling with how to write papers, give coherent talks, teach lessons in experimental psychology, learning Matlab, R and TeX and some Java, code intelligibly, or just staying flexible in supporting contingent duties in our research group VICOREG, and... stay, hm... focussed.

I have been very privileged having been able to write my dissertation as a scientific assistant in the Visual Cognition Research Group at the University of Zurich. VICOREG, as is the research group's short name, made it possible to me, not only to ameliorate my scientific knowledge and skills, but also gain lots of working experience in the field of aviation security. VICOREG would not exist today without the restless effort of my advisor Prof. Dr. Adrian Schwaninger. It is his merit, that I never run out of work during my whole time at VICOREG. He has got this intuitive way to demand from and encourage his students in an individually adaptive way, taking into account their strengths and weaknesses. He has also known when and how to give me a little push in the forward direction when I needed it.

A special thanks goes to Prof. Dr. Wolfgang Marx, who is responsible for the intellectually stimulating atmosphere at the Institute for General Psychology (Cognition) at the University of Zurich. On this occasion, I would like to congratulate Mr. Marx to his retirement and thank him for many fascinating and enthralling discussions and lectures. He is probably one of the few broad-minded personalities in the contemporary psychological research community. While ensuring highest individual freedom of his students in their detail work, he never lost sight of the big coherences, in General Psychology and hence in life.

Throughout my dissertation, I was supported by the European Commission Leonardo da Vinci Programme (VIA Project, DE/06/C/F/TH-80403, www.viaproject.eu) and especially by the Max Planck Institute for Biological Cybernetics in Tübingen for providing my fund-

ing.

Not to forget are also all national security authorities, aviation authorities, airport operators and (aviation) security companies which were willing to collaborate in our studies. Most of our studies could not have been accomplished without the aid of these public authorities and companies. Therefore many thanks go to the European Commission Leonardo da Vinci Programme, the Belgian Civil Aviation Administration, the Canadian Air Transport Authority (CATSA), the Federal Office of Civil Aviation, Switzerland, the German Ministry of the Interior, the Ministry of Justice, The Netherlands, the Amsterdam Airport Schiphol, Falck G4S, ICTS, the Berlin Airports Schönefeld and Tegel & Securitas, the Brussels Airport & Securitas, the Oslo Airport & Adecco, and most prominently Zurich Airport & Zurich State Police, Airport Division. On this occasion, I would like to thank all the thousands of professional screeners who showed the willingness and agreed to participate in our experiments. I can honestly say that I did everything to consider data privacy throughout this thesis on hand.

Finally I like to thank all the colleagues and friends I found and made on my place of employment during the last four years working at VICOREG. I always appreciated your presence and your collaboration even in times when things were frantic. I think I found friendships that will stand the test of time after my dissertation.

Dedicated to my mother Maya
and my father Toni

Chapter 1

Introduction and Summary

1.1 Structure of this thesis

This thesis falls naturally into four parts, which are highly inter-dependent. Each of the four chapters consists of a self-contained study elaborating our statistical model to eventually predict threat detection performance on the basis of image based factors. These four chapters compose the theoretical content core of this thesis.

- Introduction of the primal statistical model with a pilot study (Chapter 2)
- Model consolidation study based on a large data set (Chapter 3)
- Consolidation of the previous findings (Chapter 4)
- An example of how the statistical model proved to be valuable in the political decision-making processes (Chapter 5)

1.1.1 Chapter 2: The Model Commencements

As already mentioned, this chapter shows the first commencements in the development of a statistical model, which allows to estimate the difficulty of an x-ray image, solely based on predictor variables that can be automatically computed with the aid of image measurement algorithms. The study itself consists of four parts, each representing an experiment or an analysis. The actual experiment is a replication of an earlier study based on the X-Ray Object Recognition Test (X-Ray ORT) and is basically used to generate data to work on. This test consisted of 256 X-ray images, containing guns and knives as prohibited threat items, the participants sample consisted of 12 undergraduate students and the hit rate was used as the threat detection performance measure. The second experiment consisted in a rating experiment, where participants were asked to rate the X-Ray ORT images in terms of four image based factors. Except for minor amplifications and ameliorations these four image based factors will not significantly change over the course of this thesis. Bivariate correlations between these ratings and hit rates from Experiment 1 were all significant. In the third part we introduced a first set of mathematical/computational implementations of the image based factors. The perceptual plausibilities of these mathematical implementations were tested applying bivariate correlations between these image measurements and the human ratings. Finally, the last part of this study introduces the statistical model, applying multiple linear regression analysis to the image based factors as predictor variables to predict the hit rate. Thereby we contrasted the model using human ratings with the model using image measurements as predictor variables. The two models turned out equivalent regarding their predictive power.

1.1.2 Chapter 3: Model Consolidation I - Getting to the Bottom

As can be understood from the heading, in this chapter I report an extensive amplification of the study presented in Chapter 2. Basically, the statistical model using the image

measurements as predictor variables is put on a firm footing regarding the data set size. The X-ray image interpretation test used in this study consists of 2048 trials, and the participants sample consisted of 90 professional screeners from two European airports, as opposed to the 256 trial images and 12 undergraduate students in chapter 2. Guns, knives, improvised explosive devices (IEDs) and other, a threat item rest category, were included as threat items. Further the conducted analyses included three additional factors. Besides the original four image based factors View Difficulty, Superposition, Clutter and Opacity (equivalent to Transparency in Chapter 2) Bag Size was included in this study as a fifth image based factor. This lead to the necessity of some adaptations of the original image based factors. The other two additional factors were the human factors Training and Age. Unlike in the study reported in Chapter 2, d' was used as the main threat detection performance measure in this study. Several different types of statistical analyses were then conducted. To get a first impression of how the single factors affect threat detection performance isolated from the other factors we started with bivariate correlations between them and d'. Applying two multiple linear regression analyses separately on the image based factors and the human factors allowed us to estimate the concerted predictive power of our factors. We are very happy to present the achieved explained variances of nearly 70% for both, the image based factors regression model as well as for the human factors regression model. Furthermore the same models were applied to each threat category separately. With the model based on the image based factors we achieved an $R^2 = .60$ for guns, $R^2 = .70$ for knives, $R^2 = .34$ for IEDs and an R^2 of .77 for the threat category 'other'. With the model using the human factors Training and Age as predictor variables an $R^2 = .61$ was achieved for guns, $R^2 = .59$ for knives, $R^2 = .73$ for IEDs and an R^2 of .65 for 'other'. Here we encounter a very interesting data pattern. In short, the image measurements model prediction is clearly much weaker for IEDs than for all other categories. The exact opposite pattern can be observed with the human factors model. Finally we report an analysis of covariance (ANCOVA) which allows to investigate possible interaction effects among all our predictor variables. Results and characteristic data patterns are discussed in detail in

the Discussion section at the end of the chapter.

1.1.3 Chapter 4: Model Consolidation II - Completing the Circle

The study reported in Chapter 4 brings us back to the same structure as in Chapter 2. The four parts introduced in Chapter 2 are completed with an Experiment 5 part in this chapter. Experiments 1 - 4 are replications of the four experiments presented in Chapter 2 on a participants sample of nineteen highly experienced aviation security X-ray screening experts. Unlike in the study reported in Chapter 2 the signal detection measure d' was used as the threat detection performance measurement in this study. In contrast to the hit rate alone, d' takes into account the false alarm rate as well. Generally, the findings from Chapter 2 could be replicated almost perfectly[1] even though detection performance d' was measured for experts instead of the hit rates for unexperienced undergraduate students. Only the overall predictive power of the statistical model slightly decreased. It can be assumed, that this happened due a certain ceiling effect because the X-Ray ORT seems to be too easy for highly experienced X-ray screeners. Therefore we replicated the statistical model on image measurements in Experiment 4 again on a much larger data set in Experiment 5. The underlying X-ray image interpretation test was the same as introduced in Chapter 3, but was completed by a participants sample of 63 professional X-ray screeners from one European airport. Basically this experiment is a replication of the statistical models in Chapter 3 applied separately to each threat category. Compared to the models in Chapter 3 explained variance slightly decreased, but the characteristic data patterns remained stable, as was expected.

[1] Please note that for some image measurement formulae the direction of effect changed from Chapter 2 to Chapter 3

1.1.4 Chapter 5: Appliance in the Political Decision-Making Process

I decided to put this study at the end of the core part of this thesis to top it off. I can anticipate that the study presented in Chapter 5 built the scientific fundamentals regarding the decision-making process in the Technical Task Force of the European Civil Aviation Conference (ECAC TTF) whether a general bag size restriction should be established in Europe. The ECAC Technical Task Force holds an advisory function within the European Union. In autumn 2007 several requests have been made to introduce a new European regulation whereupon cabin baggage size should be restricted to IATA bag size[2]. We can proudly affirm that this study was the decisive factor in this political decision-making process. Based on this study the ECAC TTF has withdrawn its bag size restriction recommendation. The study was conducted on behalf of the UK Department for Transport (DfT) in collaboration with QinetiQ Ltd. In Chapter 5 I present the two core experiments of the study, both conducted by University of Zurich.

As already mentioned, the study presented in Chapter 5 consists of two independent experiments which both investigated the relative importance of our five different image based factors including Bag Size in mediating threat detection performance of human operators in airport security X-ray screening. Experiment 1 was based on a random sample of roughly 16'000 threat image projection (TIP) data records judged by approximately 700 professional X-ray screeners throughout the first half of 2007. TIP is software function available on state-of-art X-ray screening equipment that allows the projection of fictional threat items (FTIs) into the X-ray images of passenger bags during the routine baggage screening operation. χ^2-analyses revealed that the image based factors View Difficulty, Superposition and Opacity can substantially affect threat detection performance in terms of the hit rate (identification of FTIs; no false alarms could be recorded, because bags could not be controlled). Clutter and Bag Size on the other hand had no significant (negative) effect. Experiment

[2]IATA the International Air Transport Association advises passengers to travel by airplane with cabin baggages with the measures of no more than 45cm x 56cm x 25cm (IATA bag size).

2 was conducted using the offline-test introduced in the Chapters 3 and 4. Additionally, Training was included in the analyses to demonstrate the limited effectiveness of a regulation like bag size restriction compared to a quite simple intervention such as mandatory X-ray image interpretation training for professional aviation security X-ray screeners. 200 professional X-ray screeners from five European airports with varying amounts of training in X-ray image interpretation agreed in participating in this experiment. As in Chapter 3, we applied bivariate correlations, regression modeling and an analysis of covariance to the data. The results correspond very well to the results in Experiment 1. By far the largest effects were obtained with View Difficulty, Superposition and Training. Opacity showed medium effect sizes and Clutter and Bag Size showed no or very small effects, respectively. Concrete recommendations for improving X-ray image interpretation competency aviation at security check-points are given at the end of this chapter.

1.2 The whole purpose of such a statistical model and image measurements

The disposability of statistical models and automatically computable image measurements is highly valuable to estimate difficulties of X-ray images regarding threat detection performance. Statistical models on the one hand allow estimating general X-ray image interpretation difficulties. This allows taking into account differences in image difficulties in case that screener competencies are being assessed based on different stimulus material. For example, once a statistical model can explain a satisfying amount of variance in detection performance, TIP[3] data could be used to directly assess X-ray image interpretation competency without standardization. To date this is still too unfair, since TIP events differ very much in their respective difficulties. Beside the fact that the image measurements are

[3]Threat image projection (see Chapter 5 for details)

essential to build statistical models, they are very valuable for estimating screeners' individual strengths and weaknesses in dealing with the different image based factors. This fact can be used to design algorithms to guide individually adaptive training systems or TIP. State-of-the-art training systems as well as TIP systems allow to permanently record all actions and responses of the operator, well-defined by its user id, anonymous or not. Together with the possibility of controlling the stimulus material regarding our image based factors this allows to present to each operator images whose factors specifically correspond to the screeners' individual strengths and weaknesses. Learning success is known to be largest if the task difficulties are slightly challenging, i.e. the tasks should neither be too easy nor too difficult to solve successfully.

1.3 Security sensitive data

Since this thesis is based on a series of studies in collaboration with aviation security agencies, airports or security companies, who agreed in supplying us with security sensitive data for research, we highly respect and adhere to their terms and conditions regarding publication of sensitive data. Therefore I apologize that in some of the reported studies data are published in a way that no information is given that can be used to infer actual states of security of different airport or countries. In these cases detection performance mean values are not reported, airports of countries are not mentioned by name and graphs are displayed without values on the detection performance scales.

Chapter 2

The Model - Commencements

2.1 A Statistical Approach for Image Difficulty Estimation in X-Ray Screening Using Image Measurements

The relevance of aviation security has increased dramatically at the beginning of this century. One of the most important tasks is the visual inspection of passenger bags using x-ray machines. In this study, we investigated the role of image based factors on human detection of prohibited items in x-ray images. Schwaninger et al. (2004) and Schwaninger, Hardmeier, and Hofer (2005) have identified three image based factors: View Difficulty, Superposition and Bag Complexity. This article consists of 4 experiments which lead to the development of a statistical model that is able to predict image difficulty based on these image based factors. Experiment 1 is a replication of earlier findings confirming the relevance of image based factors as defined by Schwaninger, Hardmeier, and Hofer (2005) on X-ray detection performance. In Experiment 2, we found significant correlations between human ratings of image based factors and human detection performance. In Experiment 3, we introduced our image measurements and found significant correlations between them and

human detection performance. Moreover, significant correlations were found between our image measurements and corresponding human ratings, indicating high perceptual plausibility. In Experiment 4, it was shown using multiple linear regression analysis that our image measurements can predict human performance as well as human ratings can. Applications of a computational model for threat image projection systems and for adaptive computer-based training are discussed.

2.2 Introduction

The relevance of aviation security has increased dramatically in recent years and there has been substantial progress regarding screening technology, especially in the field of automatic explosive detection systems (Ying, Naidu, & Crawford, 2006). However, the last decision is always made by a human operator and investigating human factors as essential determinants of security screening performance has become an important research topic. First contributions in the field of X-ray image inspection were based on research in medical image interpretation (Gale, Mugglestone, Purdy, & McClumpha, 2000). Krupinski, Berger, Dallas, and Roehrig (2003) were able to identify important factors that influence pulmonary nodule detection. Experimental psychology studies (Ghylin, Drury, & Schwaninger, 2006) and eye movement research (McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Liu, Gale, Purdy, & Song, 2006) have been useful to better understand visual search and perceptual learning in X-ray image interpretation. A series of studies conducted in recent years has provided converging evidence for the importance of scientifically based selection, training, and testing methods to achieve and maintain high levels of performance in X-ray image interpretation (Schwaninger, 2005b, 2006b).

The aim of this study is to develop and evaluate a statistical model for image difficulty estimation in X-ray screening using image measurements. Schwaninger, Hardmeier, and

Hofer (2005) could show that there are three major image based factors which affect detection performance: View difficulty depending on the rotation of an object, Superposition by other objects in the bag, and Bag Complexity, which comprises Clutter, the bag's background texture unsteadiness, and Transparency, the relative size of dark areas in the bag. Figure 1 illustrates the three image based factors as proposed by Schwaninger, Hardmeier, and Hofer (2005).

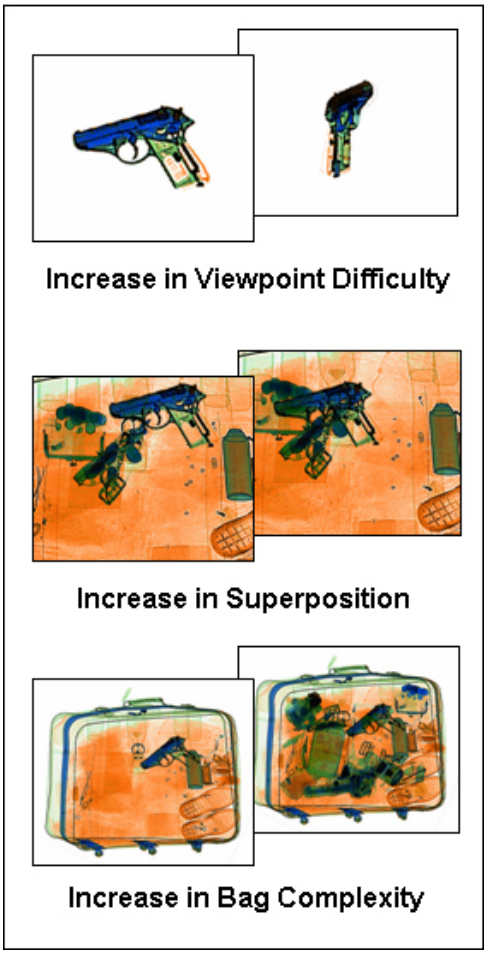

Figure 2.1: Illustration of the three major image based factors suggested by Schwaninger, Hardmeier, and Hofer (2005).

A model for image difficulty estimation using automated image measurements and human performance statistics can be very useful for threat image projection (TIP) data analysis and adaptive computer based training (CBT). TIP is a software function of state-of-the art

X-ray machines which allows the automated insertion of fictional threat items (FTIs) into X-ray images of real passenger bags. TIP systems are operational in several countries and used to enhance motivation and attention of screeners on the job. Since the TIP to bag ratio is relatively low (i.e. the number of projections per passenger bags) and the resulting TIP images (X-ray image of real passenger bag plus FTI) vary substantially with regard to image based factors, it is difficult to obtain reliable individual performance measurements. With a reliable statistical model for image difficulty estimation using image measurements, corrected individual performance scores could be calculated, which would allow more reliable individual performance assessments. A second application is adaptive computer based training. For example, the individually adaptive algorithms of X-Ray Tutor start with easy views of threat items shown in bags of low Complexity with little Superposition by other objects. Once a threat item is recognized by a screener, the View Difficulty is increased and it is shown in more complex bags with more Superposition (for details on X-Ray Tutor see (Schwaninger, 2004b)). There are large differences between individuals regarding their ability to cope with image-based factors (Schwaninger, Hardmeier, & Hofer, 2005). Therefore, a good model for image difficulty estimation using automated image measurements of image-based factors could be very useful for enhancing such individually adaptive training algorithms.

The study is sectioned into four experiments. The first experiment is a replication of earlier findings (Schwaninger, Hardmeier, & Hofer, 2005) to confirm the relevance of image based factors in predicting human performance and to show their relative inter-independence. The second experiment aims to estimate the subjective perceptual plausibility of the underlying image based factors by correlating them with the average hit rate ($p(hit)$), i.e. percent detection per image averaged across participants. Threat images were rated for View Difficulty, Superposition, Clutter, Transparency and general difficulty. Images of harmless bags were rated for Clutter, Transparency, and general difficulty. The correlation between these ratings and human detection performance reflects the relative importance of each

image based factor. We then developed statistical formulae and automated image measurements for the above mentioned image based factors. Experiment 3 was designed to estimate the perceptual plausibility of these computer generated estimates. We correlated the computer-based estimates with the corresponding human ratings to determine whether our computer-based algorithms correspond with human perception. Finally, in Experiment 4 we compared a model using computer-based estimates to a model based on human ratings of the image based factors.

2.3 Experiment 1

2.3.1 Method

Experiment 1 is a replication of the study by Schwaninger, Hardmeier, and Hofer (2005), who identified image based factors for threat item detection in X-ray image screening. Two important differences need to be mentioned. In view of possible applications in TIP systems, we are mainly interested in predicting the percentage of correct responses to images containing a threat item. Therefore, we use the hit rate instead of d' as the variable to be predicted. In our previous studies, we used the signal detection measure $d' = z(H) - z(FA)$ whereas $z(H)$ refers to the z-transformed hit rate and $z(FA)$ to the z-transformed false alarm rate (Green & Swets, 1966). Secondly, only novices and no experts are tested because we want to examine image based factors independent of expertise.

Participants

Twelve undergraduate students in psychology from the University of Zurich participated in this experiment (5 females). None of them has had any previous experience with visual inspection of X-ray images.

Materials

The X-Ray Object Recognition Test (X-Ray ORT) was used to measure detection performance. This test has been designed to analyze the influence of image based effects View Difficulty, Superposition and Bag Complexity on human detection performance when visually inspecting X-ray images of passenger bags. Inspired by signal detection theory (Green & Swets, 1966), the X-Ray ORT consists of two sets of 128 X-ray images. One set contains harmless bags without a threat item (N-trials, for noise). The other set contains the same bags, each of them with a threat (SN-trials, for signal-plus-noise). Only guns and knives of typical familiar shapes are used. This is important because the X-Ray ORT is designed to measure cognitive visual abilities to cope with effects of View Difficulty, Superposition, and Bag Complexity independent of specific visual knowledge about threat objects. The X-Ray ORT consists of 256 items (X-ray images) given by the following test design: 16 threat item exemplars (8 guns, 8 knives) x 2 View Difficulty levels x 2 Bag Complexity levels x 2 Superposition levels x 2 trial types (SN and N-trials). The construction of the items in all image based factor combinations as shown above was lead by visual plausibility criteria. After choosing two sets of X-ray images of harmless bags with different parameter values in Bag Complexity, the sixteen fictional threat items were projected into the bags in two different view difficulties at two locations with different Superposition each. The term fictional threat items (FTIs) is commonly used in connection with TIP systems as discussed in the introduction. For further details on the X-Ray ORT see (Hardmeier, Hofer, & Schwaninger, 2005; Schwaninger, Hardmeier, & Hofer, 2005). Stimuli were displayed on 17" TFT screens at a distance of about 100cm, so that the X-ray images subtended approximately 10-12 degrees of visual angle. The computer program measured outcome (hit, miss, false alarm, correct rejection) and the response times from image onset to final decision button press.

Procedure

X-ray images of passenger bags were shown for a maximum display duration of 4 seconds. Note that at airport security controls the human operators (screeners) usually have only 3-6 seconds to inspect a passenger bag. The participant's task was to decide whether the image is OK (i.e. the bag contains no threat item) or NOT OK (i.e. it contains a threat item) by clicking one of the corresponding buttons on the screen (see Figure 2.2). In addition, participants had to judge their decision confidence using a slider control (from UNSURE to SURE). These confidence ratings were used for another study. No feedback was given regarding the correctness of the responses. Participants could initiate the next trial by pressing the space bar.

Figure 2.2: Screenshot of an X-Ray ORT trial showing an X-ray image of a passenger bag containing a gun. Response buttons and slider control are aligned at the bottom of the screen.

Several practice trials were presented to make sure that the task was understood properly before the test started. Immediately prior to the actual test, all guns and knives were pre-

sented on the screen for 10 seconds, respectively. This was done to minimize any effects of threat item knowledge. Half of the items were shown in easy view and the other half in difficult view.

2.3.2 Results

Figure 2.3 displays the mean hit rate ($M = .80$) and standard deviation ($SD = 0.17$) broken up by main effects of View Difficulty, Superposition, and Bag Complexity for guns and knives. Data was first averaged across images for each participant and then across participants to calculate mean hit rates. The analysis of false alarm rates was not part of this study.

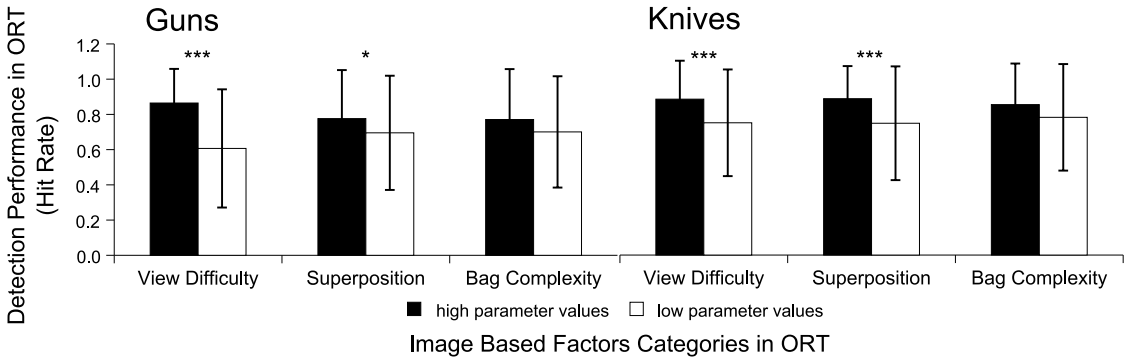

Figure 2.3: Results of Experiment 1. Mean hit rate for the detection of guns and knives, broken up by main effects of View Difficulty, Superposition, and Bag Complexity. Data was first averaged across images for each participant and then across participants to calculate mean hit rate. Error bars represent the standard deviation across participants.

Our hypothesis whereby the image based factors have great influence on threat detection performance was tested using repeated-measures ANOVA. Table 2.1 shows the effect sizes η^2, the F-statistics and the significance levels of the ANOVA main effects.

Table 2.1: ANCOVA main effects (η^2)

Guns:			
View Difficulty:	$\eta^2 = .89$	$F(1, 11) = 91.55$	$p < .001$
Superposition:	$\eta^2 = .40$	$F(1, 11) = 7.45$	$p < .05$
Bag Complexity:	$\eta^2 = .14$	$F(1, 11) = 1.76$	$p = .21$
Knives:			
View Difficulty:	$\eta^2 = .84$	$F(1, 11) = 59.06$	$p < .001$
Superposition:	$\eta^2 = .65$	$F(1, 11) = 20.48$	$p < .001$
Bag Complexity:	$\eta^2 = .23$	$F(1, 11) = 5.60$	$p = .10$

2.3.3 Discussion

We were able to replicate the results from Schwaninger, Hardmeier, and Hofer (2005) involving professional screeners fairly well regarding the main effects of View Difficulty and Superposition. However, unlike in earlier studies, the image based factor Bag Complexity had no significant effect on the hit rate for both, guns and knives. The most probable reason for this is that the threat detection performance measure used in this study was the hit rate instead of d'. As mentioned earlier, d' equals $z(H) - z(FA)$ whereas $z(H)$ refers to the z-transformed hit rate and $z(FA)$ to the the z-transformed false alarm rate (Green & Swets, 1966). Effects of Bag Complexity are more likely to be found on the false alarm rate. In X-ray screening tests, the false alarm rate is based on the number of times a participant judges a bag to be NOT OK even though there is no threat item in it. Consistent with this view, we found clear effects of Bag Complexity on d' in earlier studies (Hardmeier et al., 2005; Schwaninger, Hardmeier, & Hofer, 2005). It is therefore not too surprising that we could not find a significant effect of Bag Complexity on hit rate alone in Experiment 1.

2.4 Experiment 2

Experiment 2 was designed to investigate the perceptual plausibility of our image measurements introduced in Experiment 3.

2.4.1 Method

The same students who had participated in Experiment 1 took part in Experiment 2 one week later. The participant's task was to rate the difficulties of the X-Ray ORT images regarding View Difficulty and Superposition of the threat images. In addition, Clutter, Transparency and general item difficulty had to be rated for threat and non-threat images. The ratings were given by mouse clicks on a 50-point scale (0 = very low to 50 = very high). No initial position was set. Figure 2.4 shows a screenshot of an bag containing a threat item.

2.4.2 Results

In order to estimate the relative importance of image based factors (Schwaninger, Hardmeier, & Hofer, 2005) on human detection performance, we correlated ratings for View Difficulty, Superposition, Clutter and Transparency (Experiment 2) with the hit rates obtained in Experiment 1. Data analysis was conducted separately for guns and knives.
Figure 2.5 shows the averaged ratings across all participants and across all threat items. The ordinate depicts the rating scores on the 50-point scale (see Figure 2.4). The black and white bars in each image based factors category represent the low and high parameter values according to the arrangement of the X-Ray ORT test design. Over-all mean rating value was $M = 19.2$ with a standard deviation of $SD = 15.4$. Inter-rater consistency was quite high with an average correlation (Fisher-corrected) between subjects of $r = .64$ for View

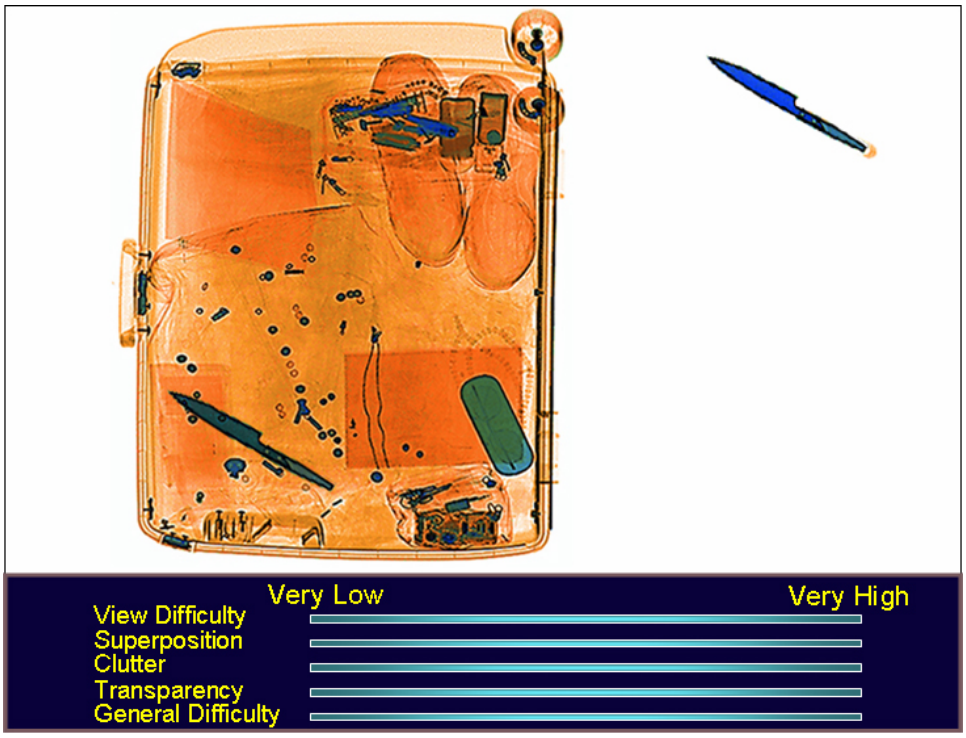

Figure 2.4: Screenshot of a typical trial of Experiment 2 containing a knife. All participants were asked to judge the image based factors subjectively, whereby Bag Complexity is separated in Clutter and Transparency. Additionally, participants were asked to judge the general item difficulty as well (not analyzed in this study). Threat items were displayed next to the bag. For non-threat items, the slider controls for View Difficulty and Superposition were discarded.

Difficulty, $r = .62$ for Superposition, $r = .65$ for Clutter and $r = .40$ for Transparency.

Correlations of ratings of image based factors with hit rate per image averaged across the participants of Experiment 1. Table 2.2 shows the correlations between the human ratings of the image based factors and the hit rates from Experiment 1. Image based factors and hit rates were averaged across participants.

Concerning the mathematical signs, note that the hit rate points in the opposite direction of threat detection difficulty. The more difficult a threat item is to be detected the lower the hit rate.

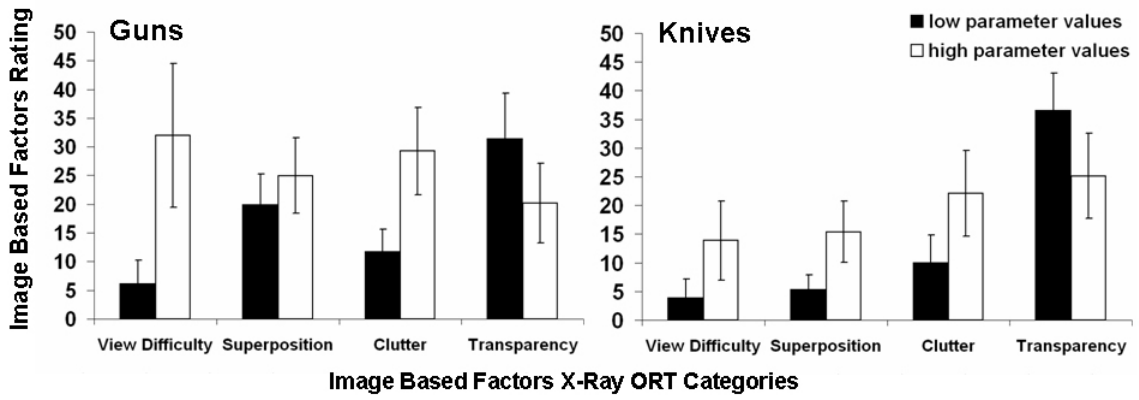

Figure 2.5: Results from Experiment 2 for guns and knives separately. The image based factor Bag Complexity from the X-Ray ORT is split into the sub-factors Clutter and Transparency according to the rating experiment design shown in Figure 2.4. Please note that the factor Transparency points in the opposite direction compared to Bag Complexity and the other image based factors.

2.4.3 Discussion

All subjective human ratings show significant correlations with the hit rates from Experiment 1, except for Clutter in X-ray images containing a knife, which was marginally not significant ($p = .06$). Thus, the results from Experiment 1 and Experiment 2 showed that image based factors affect objective X-ray image difficulty (hit rate) and the image-based factors can be rated by novices. Consistent with the findings from Experiment 1, the ratings of image based factors show that Clutter and Transparency are less predictive than ratings of View Difficulty and Superposition. For the development of image measurements, it was necessary to split up the factor Bag Complexity into Clutter and Transparency. However, this seems to be problematic, because for subjective ratings they seem to be highly interdependent. The ratings of Clutter and Transparency are highly correlated: $r(12) = -.93, p < .001$ for guns and $r(12) = -.86, p < .001$ for knives. We return to this issue in Discussion section.

Table 2.2: Correlations between image based factors and $p(hit)$

Guns:		
View Difficulty:	$r(12) = -.56$	$p < .001$
Superposition:	$r(12) = -.69$	$p < .001$
Clutter:	$r(12) = -.32$	$p < .05$
Transparency:	$r(12) = .37$	$p < .01$
Knives:		
View Difficulty:	$r(12) = -.53$	$p < .001$
Superposition:	$r(12) = -.67$	$p < .001$
Clutter:	$r(12) = -.24$	$p = .06$
Transparency:	$r(12) = .31$	$p < .05$

2.5 Experiment 3

The aim of Experiment 3 was to develop computer-based algorithms to automatically estimate the image based factors View Difficulty, Superposition, Clutter, and Transparency. The perceptual plausibility of these computer-based algorithms was examined by correlating them with the human ratings obtained in Experiment 2.

2.5.1 Method

All image measurements developed for this purpose are based on theoretical considerations. Different algorithm parameters were optimized by maximizing the correlations between the image-based factors estimates and detection performance measures derived from earlier X-Ray ORT findings from X-ray screening experts.

Statistical estimates and image measurements for image based factors

View Difficulty

Even with the aid of 3D volumetric models, it is not (yet) possible to satisfyingly determine the degree of a 3-dimensional rotation (View Difficulty) of a physical threat item automatically from its 2-dimensional X-ray image (Mahfouz, Hoff, Komistek, & Dennis, 2005). Additional difficulties regarding image segmentation arise from the very heterogeneous backgrounds of X-ray images, compare (Sluser & Paranjape, 1999). Therefore, this image based factor is not (yet) being calculated by image processing, but statistically from X-Ray ORT detection performance data obtained in Experiment 1.

$$VD_j = \frac{\left(\sum_{i=1}^{4} \text{HitR}_i\right) - \text{HitR}_j}{3} \tag{2.1}$$

Equation 2.1 shows the calculation of the image based factor View Difficulty, whereas i is the summation index ranging from 1 to 4 (2 bag complexities x 2 Superpositions), j denotes the index number of the X-ray image in question (one threat exemplar in one of the two views), $HitR_j$ is its average hit rate across all participants and '4' is the number of the bags each FTI was projected into. In order to avoid a circular argument in the statistical model (multiple linear regression, see Experiment 4) by partial inclusion of the criterion variable into a predictor, the hit rate of the one item in question is excluded from this estimate.

It is important to understand that this concept of View Difficulty is not just reflecting the degree of rotation of an object. In that case there would be two parameter values for all threat exemplars only. View difficulty as it is conceptualized here reflects innate View Difficulty attributes unique to each exemplar view separately.

Superposition

This image based factor refers to how much the pixel intensities at the location of the FTI

in the threat bag image differ from the pixel intensities at the same location in the same bag without the FTI. Equation 2.2 shows the image measurement formula for Superposition. $I_{SN}(x, y)$ denotes the pixel intensities of a threat image and $I_N(x, y)$ denotes the pixel intensities of the corresponding harmless bag.

$$SP = \sqrt{\sum_{x,y} \left(I_{SN}(x, y) - I_N(x, y)\right)^2} \tag{2.2}$$

It should be noted that this mathematical definition of Superposition is dependent on the size of the threat item in the bag. For further development of the computational model it is conceivable to split up Superposition and the size of the threat item into two separate image based factors. Measurement of Superposition would require having both the bag with the FTI and without. For both applications mentioned in the introduction, this is possible with current TIP and CBT technology. In TIP, the FTI, its location, the bag with and without the FTI are recorded. In state-of-the-art computer-based training systems, the same information is recorded and stored, too.

Clutter

This image based factor is designed to express bag item properties like its textural unsteadiness, disarrangement, chaos or just Clutter. In terms of the bag images presented, this factor is closely related to the amount of items in the bag as well as to their structures in terms of complexity and fineness. The method used in this study is based on the assumption, that such texture unsteadiness can be described mathematically in terms of the amount of high frequency regions.

$$\text{CL} = \sum_{x,y} I_{hp}(x, y) \tag{2.3}$$

$$\begin{aligned} \text{where} \quad & I_{hp}(x, y) = I_N * \mathcal{F}^{-1}(hp(f_x, f_y)) \\ \text{and} \quad & hp(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{f_x^2 + f_y^2}}{f}\right)^d} \end{aligned}$$

Equation 2.3 shows the image measurement formula for Clutter. It represents a convolution of the empty bag image (N for noise) with the convolution kernel derived from a high-pass filter in the Fourier space. I_N denotes the pixel intensities of the harmless bag image. $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. $hp(f_x, f_y)$ represents a high-pass filter in the Fourier space (see Appendix A).

Transparency

The image based factor Transparency reflects the extent to which X-rays are able to penetrate objects in a bag. This depends on the specific material densities of these objects. These attributes are represented in X-ray images as different degrees of luminosity. Heavy metallic materials such as lead are known to be very hard to be penetrated by X-rays and therefore appear as dark areas on the X-ray images.

$$TR = \frac{\sum_{x,y}(I_N(x, y) < \text{threshold})}{\sum_{x,y}(I_N(x, y) < 255)} \tag{2.4}$$

Equation 2.4 shows the image measurement formula for Transparency. $I_N(x, y)$ denotes the pixel intensities of the harmless bag. $threshold$ is the pixel intensity threshold beneath which the pixels are counted. The implementation of the image measurement for the image based factor Transparency is simply achieved by counting the number of pixels being darker than a certain threshold (e.g. < 65) relative to the bag's overall size (< 255, non-white pixels).

2.5.2 Results

To examine perceptual plausibility of the computer-based measurements, we correlated them with the corresponding averaged ratings from Experiment 2. Table 2.3 shows the Pearson's product-moment correlations between the calculated measurements and the corresponding human ratings' mean values for each image based factor and threat category separately.

Table 2.3: Correlations between image based factors and human ratings

Guns:		
View Difficulty:	$r(12) = -.62$	$p < .001$
Superposition:	$r(12) = -.54$	$p < .001$
Clutter:	$r(12) = .16$	$p = .20$
Transparency:	$r(12) = -.69$	$p < .001$
Knives:		
View Difficulty:	$r(12) = -.47$	$p < .001$
Superposition:	$r(12) = -.44$	$p < .001$
Clutter:	$r(12) = .18$	$p = .16$
Transparency:	$r(12) = -.63$	$p < .001$

2.5.3 Discussion

Except for Clutter all correlations between automated measurements and ratings are highly significant. In the discussion of Experiment 2 the high inter-correlations between the human ratings of the image based factors Clutter and Transparency was mentioned ($r(12) = -.93, p < .001$ for guns and $r(12) = -.86, p < .001$ for knives). Consistent with this

result, there were also fairly high inter-correlations between the corresponding calculated estimates of the image based factors Clutter and Transparency ($r(64) = .52, p < .001$ for guns and $r(64) = .55, p < .001$ for knives). Except for Clutter, we can conclude that our algorithms for automated estimation of image based factors are perceptually plausible because they correlate significantly with the ratings of novices.

2.6 Experiment 4

Experiment 4 was designed to evaluate the predictive power of a statistical model based on automated estimation of image based factors. To this end, we now compare the results of multiple linear regression analysis using the automated estimates of image based factors as predictors with the results of multiple linear regression analysis using the human ratings of image based factors as predictors.

2.6.1 Method

The comparison included the four image based factors introduced in Experiment 3.

Multiple Linear Regression Analysis

The predictors of the multiple linear regression model are our image based factors; the hit rates per image averaged across participants (Experiment 1) is the dependent variable. We compared the two statistical models in terms of their goodness-of-fit measures, their regression coefficient's significances and the percentage of variance in the dependent variable hit rate the model is able to explain by its predictors.

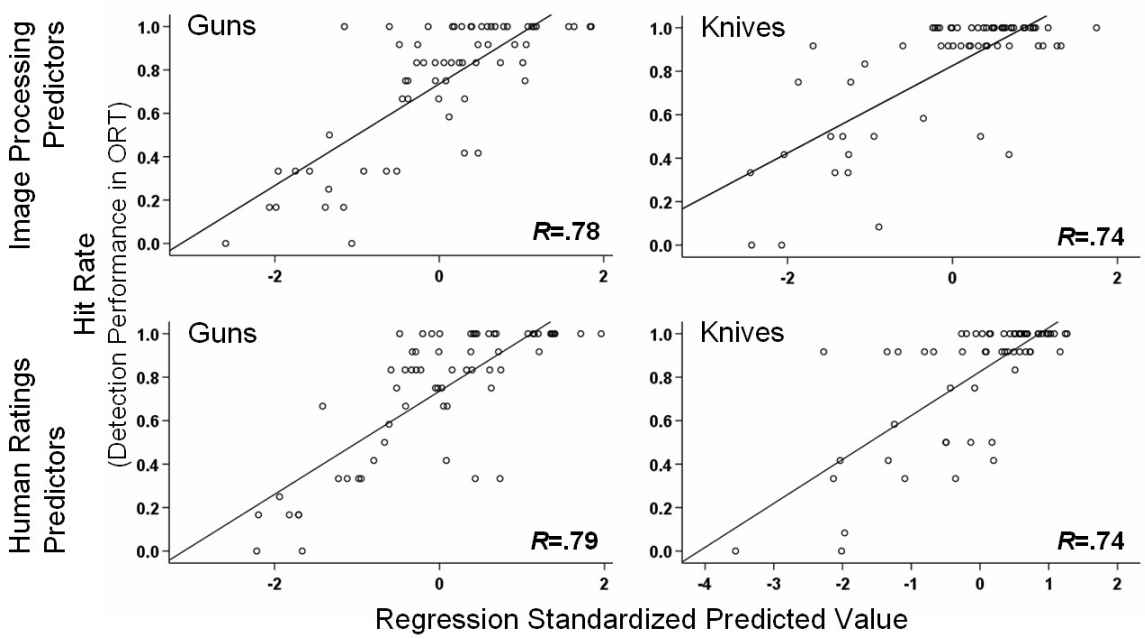

Figure 2.6: The four scatter plots from the models predicting the hit rate on the basis of all disposable image based factors as predictors. Guns and knives are displayed separately (columns). The models based on the calculated predictors derived from image measurements are displayed in the first row and the models based on rated image based factors predictors are displayed in the second row.

2.6.2 Results

Figure 2.6 shows the scatter plots with regression standardized predicted values on the abscissa and the actually measured hit rate from Experiment 1 on the ordinate.

Table 2.4 shows the most important statistical values of the four multiple linear regression analyses arranged in columns and rows like in Figure 2.6. The single tables show the four predictors in the rows. The first column gives the variable names of the image based factors. Standardized regression weights for guns are given in the second column and the third column shows the p-value statistics indicating the significance of the single regression coefficients in the model. This recurs in columns four and five for knives. For both models, based image measurements and human ratings, also the model summaries for guns and

knives are given. R^2 tells us to which extent the model is able to predict the variance in the hit rate. Because R^2 increases with the number of predictors independently of their predictive power, $R^2(adj)$ taking into account the number of predictors, is given too. Finally the statistical indices F-value and the significance level of the model as a whole (p-value) are given.

All statistical models are highly significant in the overall goodness-of-fit verification statistics, both for guns and knives. The R^2-values, the extent to which a model is able to explain the variance in the dependent variable by its predictors, are very high compared to values usually obtained when predicting human performance. The model based on our image measurements achieves an R^2 of .61 ($R^2(adj)$=.58) with guns and an R^2 of .54 ($R^2(adj)$=.51) with knives. The ratings model is even marginally better with an R^2 of .62 ($R^2(adj)$=.60) with guns and an R^2 of .55 ($R^2(adj)$=.52) with knives. Concerning the regression coefficients in detail, the predictors View Difficulty and Superposition are always significant, mostly highly significant. This is not the case for the two sub-factors of Bag Complexity (Clutter and Transparency).

2.6.3 Discussion

The different statistical models in Experiment 4 show that the image based factors suggested by Schwaninger, Hardmeier, and Hofer (2005) are quite powerful predictors of human detection performance. The model based on automated estimation of image-based factors is as predictive as human ratings. Admittedly, Experiment 4 shows also that the sub-factors of the image based factor Bag Complexity, Clutter and Transparency, do not contribute significantly to the explanatory power of the model. In some cases, they even show regression weights which point in the opposite direction of what is expected. As already mentioned in Experiments 1 and 2 this can be explained by the fact that in detection experiments Bag Complexity rather affects the false alarm rate than the hit rate.

Model Summaries (Per Category)					
		Guns		Knives	
		β -weights	p-values	β -weights	p-values
Image Measurements	View Difficulty	.61	.000	.29	.018
	Superposition	.33	.000	.50	.000
	Clutter	.11	.239	-.15	.194
	Transparency	-.18	.072	-.03	.788
	R^2	.61		.54	
	$R^2(adj)$	.58		.51	
	$F(4, 59)$	23		17	
	p	.000		.000	
Human Ratings	View Difficulty	-.39	.000	-.33	.001
	Superposition	-.60	.000	-.68	.000
	Clutter	.36	.227	-.13	.497
	Transparency	.27	.107	-.34	.123
	R^2	.62		.55	
	$R^2(adj)$	.60		.52	
	$F(4, 59)$	24		18	
	p	.000		.000	

Table 2.4: Statistical analysis tables of the models with the most important statistical values of the multiple linear regression analyses. Each of the four tables shows the statistical values of the verification of each regression coefficient separately in the rows. Additionally the model's overall goodness-of-fit verification values are given in the bottom row of each model. In both statistical models, the dependent variable is the hit rate obtained in Experiment 1.

This is currently being investigated in additional experiments. Nevertheless, our computational model is able to predict the hit rate in terms of image based factors as good as human ratings can. Such a model could therefore provide a basis for the enhancement of individually adaptive computer-based testing and training systems in which the estimation of X-ray image difficulty is an essential component. In addition, the image measurements developed in this study can be very useful for analyzing more reliable individual TIP performance scores by taking into account image difficulty as explained in the introduction. It is interesting to discuss the differences in the beta-weights between guns and knives in the image processing model. For guns View Difficulty is weighted almost double compared to Superposition. For knives, where Superposition is weighted almost double compared to View Difficulty, the contrary pattern was observed. We are currently conducting additional analyses to find out whether this effect is related to differential changes by 3D rotation. The reason why Superposition is weighted much stronger in knives than in guns is probably due to the Superposition formula which also reflects the size of the threat items. In the X-Ray ORT knives differ more in size than guns. Thus, the regression coefficient patterns reflect actual characteristics of the weapon categories. The scatter plots (Figure 2.6) reveal that, especially in knives, there is a certain ceiling effect. Therefore, it might be of value to use non-linear regression for modeling hit rates in the future. Apart from that, this study can be viewed as the basis for further statistical models for the prediction of individual screener responses to single X-ray images using binary logistic regression. In addition, together with the development of enhanced and additional image based predictors we intend to develop parallel statistical models to predict hit rates as well as false alarm rates.

Chapter 3

Model Consolidation I - Getting to the Bottom

3.1 On How Image Based Factors and Human Factors Contribute to Threat Detection Performance in X-Ray Aviation Security Screening

Schwaninger, Michel, and Bolfing (2007) contributed an article on X-ray image difficulty estimation based on a set of image based factors. The study revealed that it is possible to predict average detection performance (across a sample of participants) on a single image quite well solely based on computationally accessible image properties. The image based factors used in that model were View Difficulty, Superposition, Clutter and Transparency. All image based factors can be automatically calculated. Multiple linear regression was used for statistical modeling. A comparison between the model based on automatically computed predictors (image based factors) and the same model based on human ratings (of the image based factors) revealed that our image measurements and statistics can predict

human performance as well as human raters can. The study was based on a participants sample of 12 undergraduate students and on a X-ray object recognition test consisting of 256 images.

The study reported in the following is an extensive amplification of the Schwaninger, Michel, and Bolfing (2007) article. We were able to replicate the results of the earlier study with professional screeners and additional extensions in terms of data set size, additional factors and additional statistical analyses. The new test consists of 2048 test items and results are based on a participants sample of 90 screeners. Furthermore, three additional factors have been included: Human factors, namely Training and Age as well as the image based factor Bag Size. The number of threat categories was doubled by adding improvised explosive devices (IEDs) and 'other' to guns and knives.

Since detection performance in visual search tasks depends on the stimulus material (image based factors) but also on human factors, human factors should not be neglected in a comprehensive model. This allows comparing the impact of image based factors and human factors. The image based factors included are Threat Category, View Difficulty, Superposition, Opacity, Clutter, and finally Bag Size as a new factor. Starting with a summary of bivariate correlations between all factors and detection performance d' we give a first impression of the relationship between the predictors and detection performance. Subsequently two multiple linear regression analyses with image based factors and human factors as predictors are presented. These models allow an estimation of the total amount of variance in d' explained by the image based factors and human factors, respectively. Finally, an analysis of covariance (ANCOVA) is reported to reveal interactions between the factors. It allows estimating the main effects as well as their interactions.

3.2 Introduction

Together with the tremendous growth of civil aviation the importance of aviation security and its public perception has dramatically increased in the last few decades ("Annual Review of Civil Aviation 2005", 2006). The security checkpoints at the gates for X-raying passenger bags are the key element in aviation security all over the world. Despite great improvements in technical equipment, including high resolution X-ray machines, the decision whether a piece of luggage can enter an airplane or not is still made by human screeners. Therefore aviation security officers and their activity in screening passenger bags are a critical link of utmost importance in aviation security. In this study we analyze the effects on detection performance of prohibited items in passenger bags of two different groups of factors: 'Human factors' and 'image based factors'. The concept of image based factors subsumes all properties of the passenger bags' X-ray images that are relevant in mediating performance in detecting prohibited items. The concept of human factors subsumes available properties of the persons performing the screening task relevant in mediating threat detection performance. The aim of the reported study in this article is to investigate the role of image based and human factors on the threat detection performance in passenger bag screening tasks. For this purpose the effects of the different factors on threat detection performance, as well as their interactions will be assessed as comprehensively as possible. Previous work (Schwaninger, 2003b; Schwaninger, Hardmeier, & Hofer, 2005; Schwaninger, Michel, & Bolfing, 2007) has identified the following performance relevant image based factors: Threat Object View Difficulty, Superposition by other objects and Bag Complexity (represented in the following by Opacity and Clutter). The experiment is based on an off-line computer based test consisting of 2048 trials. The test is designed with the four image based factors View Difficulty, Superposition, Bag Complexity and Bag Size systematically varied in order to avoid confounded variables. This design allows analysis of individual and combined effects of the image based factors, as well as analysis of their interactions. Furthermore we will analyze data of the human factors Training and

Age (Riegelnig & Schwaninger, 2006). Training was operationalized as the amount of hours spent on training using the 'X-Ray Tutor' computer based training system prior to testing.

3.2.1 Image Based Factors

Schwaninger (2003b) and Schwaninger, Hardmeier, and Hofer (2005) have identified three image based factors which affect human threat detection performance significantly: View Difficulty, Superposition, and Bag Complexity. These image based factors have been modeled into mathematical formulae (Schwaninger, Michel, & Bolfing, 2007; Bolfing & Schwaninger, 2007). View Difficulty is implemented as an a posteriori calculable value named FTI View Difficulty. The abbreviation FTI represents fictional threat item, X-ray images of threat objects being artificially projected into X-ray images of passenger bags. Superposition and Bag Complexity are implemented as image processing measurements whereby Bag Complexity is split up into Clutter and Opacity. The introduction of the image based factor Bag Size in this study necessitated normalization of earlier implementations of Clutter and Opacity regarding bag size.

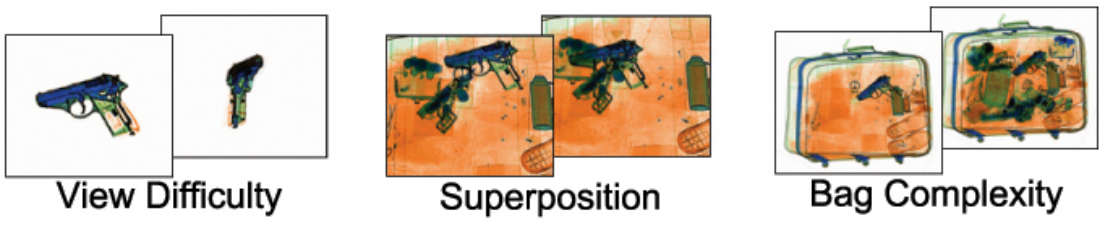

Figure 3.1: Image based factors

FTI View Difficulty

The general formula for FTI View Difficulty can be seen in Equation 3.1. It is a slight modification of the mean of the inverted detection performance value (DetPerf) over all items (index N_{OV}) containing the same FTI object (subindex O) in the same view (subindex V) as the item in question. 'Inverted' refers to the fact that the measured detection performance is subtracted from a theoretical maximum detection performance (max(DetPerf)), in order to ensure that high values of FTI View Difficulty correspond with high difficulties. The slight modification refers to the exclusion of the item in question from averaging.

$$\text{FtiVD}_{OVj} = \frac{\sum_{i=1,j\neq i}^{N_{OV}} (max(\text{DetPerf}) - \text{DetPerf}_{OVi})}{N_{OV} - 1} \tag{3.1}$$

Superposition

Superposition is modeled as the inverted Euclidean distance in pixel intensity between an SN image (signal plus noise; image containing a threat item) and its corresponding N image (noise; non-threat image).

$$SP = C - \sqrt{\sum_{x,y} \left(I_{SN}(x,y) - I_N(x,y)\right)^2} \tag{3.2}$$

Please note that in all reported analyses we used logarithmically transformed Superposition values. After inspecting the scatterplots of all our factors with the detection performance d' in order to check for non-linear relationships (a violation of the requirements of multiple linear regressions) we decided to linearize Superposition by this log-transform. This way heteroscedasticity (another violation of the requirements of MLR) can be avoided and the

explained variance of the relationship between the factor and detection performance can be increased.

Clutter

Clutter is designed to capture bag image properties like disarrangement, textural noise, chaos or just plain clutter. We modeled the Clutter variable based on the assumption that image properties like the ones mentioned above correspond with larger amounts of high spatial frequencies in the image. Equation 3.3 represents a convolution of the empty bag image (N for noise) with the convolution kernel derived from a high-pass filter in the Fourier space. I_N denotes the pixel intensities of the harmless bag image. $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. $hp(f_x, f_y)$ represents a high-pass filter in the Fourier space. BS represents Bag Size (see Equation 3.5). Cut-off frequency f and transition d (the filter's order) were set to $f = 0.03$ and $d = 11$. The pixel summation on the high-pass filtered image was restricted to the bag's area.

$$\text{CL} = \frac{\sum_{x,y} I_{hp}(x,y)}{BS} \tag{3.3}$$

$$\begin{aligned} \text{where} \quad I_{hp}(x,y) &= I_N * \mathcal{F}^{-1}(hp(f_x, f_y)) \\ &= \mathcal{F}^{-1}(\mathcal{F}(I_N \cdot hp(f_x, f_y)) \\ \text{and} \quad hp(f_x, f_y) &= 1 - \frac{1}{1 + \left(\frac{\sqrt{f_x^2 + f_y^2}}{f}\right)^d} \end{aligned}$$

Opacity

Opacity represents how well X-rays are able to penetrate an object. High Opacity values represent low penetrability. In X-ray images this property is represented by pixel color and

brightness. Opacity represents the total size of areas with pixels being darker than a certain threshold relative to the bag's size. In Equation 3.4 all pixels being darker than a certain threshold (e.g. 64) are summed up and divided by the bag's size (Bag Size as denominator).

$$\text{OP} = \frac{\sum_{x,y} \big(I_N(x,y) < 64\big)}{\text{BS}} \tag{3.4}$$

Bag Size

The Bag Size formula below is applicable to grayscale images with pixel brightness values ranging from 0 (black) to 255 (white). All pixels with brightness values lower than 254 (almost white) are considered as part of the bag. Bag Size is then defined as the size of the bag in number of pixels

$$\text{BS} = \sum_{x,y} (I_N(x,y) < 254) \tag{3.5}$$

3.3 Methods and Procedures

3.3.1 Participants

The participants sample consists of 90 professional aviation security X-ray screening officers from two European airports (48 females). On average females are 40.6 and males 35.9 years old with standard deviations of 17.8 and 13.6 years respectively.

3.3.2 Stimuli

The 2048 test stimuli were created automatically using the image measurement algorithms described above. The number of trial images is determined by the following test design: The test consists of eight threat exemplar pairs per threat category. Given the categories Guns, Knives, IEDs and Other this results in 64 different exemplars of threat items. Each of these threat items is presented with each possible factor combination. Each of the image based factors introduced above is implemented in the design with two dichotomous parameter values representing low and high values. For View Difficulty, Superposition, Bag Complexity and Bag Size this results in 2x2x2x2= 16 factor combinations. The 64 threat exemplars in 16 factor combinations result in 1024 images. In order to apply signal detection theory (Green & Swets, 1966) in the analysis all 1024 bag images containing fictional threat items (FTIs) are also presented in the test not containing any threat items. This results in the total of 2048 images.

The construction process of the test stimuli was partly manual and partly automated. In a first step the 64 threat exemplars were chosen manually such that the diversity of threat items in each of the four categories is well represented. In a second step a set of 1024 bag images was chosen based on the image measurements introduced above. In total 6659 bag images were analyzed regarding Clutter, Opacity and Bag Size. Subsequently we determined the membership of each image regarding high or low parameter values by applying median splits on each of the three image based factor distributions Opacity, Clutter and Bag Size. Opacity and Clutter are very highly intercorrelated. Thus it did not make sense to define and vary high and low parameter values for Opacity and Clutter independently. Instead the dichotomous variable Bag Complexity was defined based on Opacity and Clutter: For Bag Complexity high and low parameter values were defined as bags with both high or low Opacity and Clutter values, respectively. Bags with high Opacity and low Clutter values or vice versa were discarded. For each of the resulting factor combinations - low/high Bag Complexity x small/large Bag Size - 256 images were chosen manually. In the last

step fictional threat items were automatically merged with the harmless bags. Each of the 64 threat exemplars exists in two different views. The easy views were depicted in frontal view and the difficult ones were depicted in a 85° rotation relative to the frontal view, either horizontally or vertically. This results in a total of 128 fictional threat items (FTIs). Each of these 128 FTIs was finally merged with the 256 harmless bags with two different levels of Superposition - low and high. This procedure was applied four times for each combination of harmless bags. As already mentioned, this process was fully automatic. The underlying merging algorithm merges the images and calculates the Superposition value. If the Superposition value lies in the low or high superposition level range it is being saved as such. If not, the process is repeated until the FTI can be merged within the desired superposition value range.

3.3.3 Procedure

Since the test contains a very large amount of items participants completed it over multiple sessions. The test presentation was implemented within the computer based training system X-Ray Tutor 2.0 which can be run as a testing and a training environment. X-Ray Tutor is a well-established training tool designed to effectively improve and reliably test X-ray image interpretation competency. Customer airports are recommended to advise their security screeners to practice at least 20 minutes per week using X-Ray Tutor. The current test was inserted into the familiar training sequences and continued after each login until completion. After that, normal training continued. The shared basis of training and testing allows extracting training data of each screener prior to testing.

The 2048 images are presented in random order. The participants' task is to decide whether a piece of luggage would be OK or not OK to hypothetically enter an airplane by pressing buttons OK or NOT OK with a computer mouse. Unlike in training mode, in testing mode participants receive no feedback with regards to the correctness of their answers. Data are

recorded in a database.

3.3.4 Statistics

The data are analyzed in three ways. First we present all bivariate correlations between the independent variables (image based factors and human factors) and detection performance d'. This gives a good estimation of how well detection performance can be predicted on the basis of our predictors. The second statistics we present are two separate linear regression models, one for image based factors and one for human factors, respectively. Regression analyses allow estimating the combined impact of the respective predictors together. Regression analyses allow estimating to what extent a certain set of predictors is able to predict the measured values (Coolican, 2004), in this case the measured detection performance d'. The regression analysis using the image based factors as predictors is a replication and refinement of an earlier study by Schwaninger, Michel, and Bolfing (2007) which was based on the X-Ray ORT (Hardmeier, Hofer, & Schwaninger, 2006b). The present analysis is based on a much larger item- and subject sample since we used a new test for the present study. The most notable differences between the X-Ray ORT and this test are the inclusion of the new image based factor Bag Size and the extension of the threat item categories by IEDs (Improvised Explosive Devices) and "Other". Since linear regression analyses do not take into account any kind of interaction effects between the predictors we report a third type of analysis. The analysis of covariance (ANCOVA), with our image based factors operationalized as repeated measures variables and the human factors as covariates. It is important to be aware of potential interaction effects, since the presence of large interactions would limit the amount of variance explained by the multiple linear regression models.

3.4 Results

In conformity with the Statistics section the results are reported in the following order: First we report the findings of the bivariate and partial correlations. In a second step we report the results of the multiple linear regression analyses per threat item category as well as for all categories combined. One set of multiple linear regression analyses will be based on image based factors and the other on human factors. Finally we report the ANCOVA with the image based factors as repeated measures and with the human factors Age and Training as covariates.

3.4.1 Bivariate Correlations

Figure 3.2 shows the bivariate correlations with d' of each image based factor and the partial correlations with d' of each human factor - with the respective other human factor serving as a control variable. The reason why we decided to treat image based factors and human factors differently is the following: The image based factors have been implemented within the computer based test in order to obtain orthogonal relationships between them. In other words, image based factor values vary independently across test items. Since we could not ensure independence of the human factors Age and Training through test design or sample selection, orthogonal relationships between human factors cannot be assumed. The data reveal that indeed Age and Training are confounded, with people tending to train more the older they are. Therefore we decided to additionally report partial correlations to avoid false conclusions regarding the effects of Age and Training on visual search tasks. Furthermore we decided to graphically report R^2 values instead of plain Rs. The great advantage of R^2 over R is that it can be directly interpreted as the amount of variance in the dependent variables (d') that can be explained by the independent variable (single factors). The disadvantage is the loss of information on the sign of the relationship due to squaring.

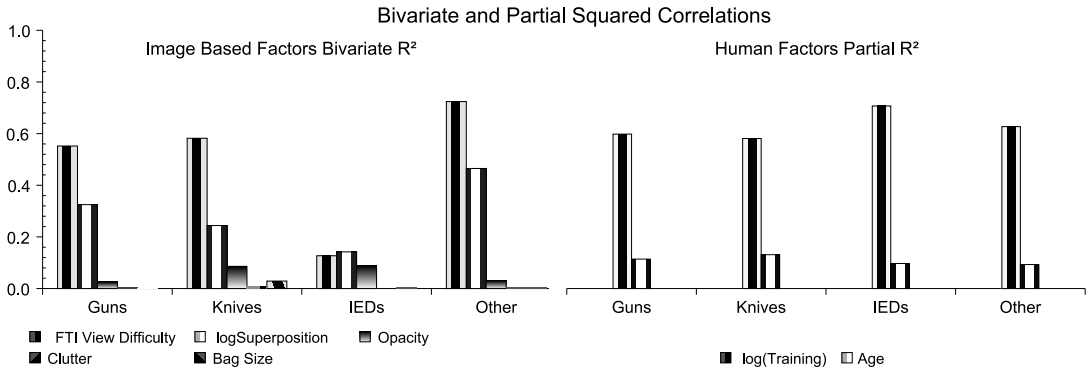

Figure 3.2: The bivariate R^2 values for the image based factors and the partial R^2 values for the human factors are estimates for the amount of variance in detection performance d' that can be explained by the single factors.

Figure 3.2 illustrates the relationship of the individual factors on detection performance d' by threat category for image based factors and human factors separately. The graphs clearly reveal that three factors explain a substantially higher amount of variance than all the others, i. e. FTI View Difficulty, Superposition (log-transformed) and Training (log-transformed training hours). Age also shows a notable effect which is much smaller, but remains stable across all threat categories. Exact values are reported in Table 3.1. A detailed discussion on the data patterns is given in the Discussions section at the end of this article.

3.4.2 Multiple Linear Regression Analysis

Figures 3.3 - 3.5 all show scatterplots illustrating the statistical relationship between the observed (empirically measured) detection performance values d' (ordinate) and the standardized predicted values estimated by the respective multiple linear regression models (abscissa). For each model R^2 and R values are displayed in the bottom right corner of the scatterplot as a measure for the closeness of the relationship between model prediction and empirical measurements.

	Image based factors bivariate correlations with d'					**Human factors** partial correlations with d'		**Human factors** bivariate correlations with d'	
	FTI-VD	logSP	OP	CL	BS	logTR	Age	logTR	Age
Guns	-.74	-.57	-.16	-.06	-.02	.77	-.34	.75	.19
Knives	-.76	-.49	-.29	-.09	-.17	.76	-.36	.73	.16
IEDs	-.36	-.38	-.30	-.02	-.05	.84	-.31	.84	.28
Other	-.85	-.68	-.18	-.05	-.05	.79	-.31	.78	.24

Table 3.1: Tabulation of the correlations between the single factors (human and image based) and the detection performance measure d' separately for each threat category. For human factors, also partial correlations are given, the respective other human factor taken as the control variable.

Category independent models

Figure 3.3 shows the scatterplots of the multiple linear regression models for the image based factors and human factors respectively. Differences concerning threat categories are not taken into account here. Both models can explain nearly 70% of the observed variance in d'. In the image based factors model 1024 data points are estimated. Each data point represents one signal-noise/noise item pair with its hit rate and false alarm across all 90 screeners. In the human factors model there are only 90 data points because d' values are calculated per subject across all 1024 item pairs.

Table 3.2 shows the most important statistical values of the multiple linear regression analyses for both models.

Models by category

Figures 3.4 and 3.5 show the corresponding linear regression model scatterplots to Figure 3.3 but separately for each threat category. This allows us to compare the relationship

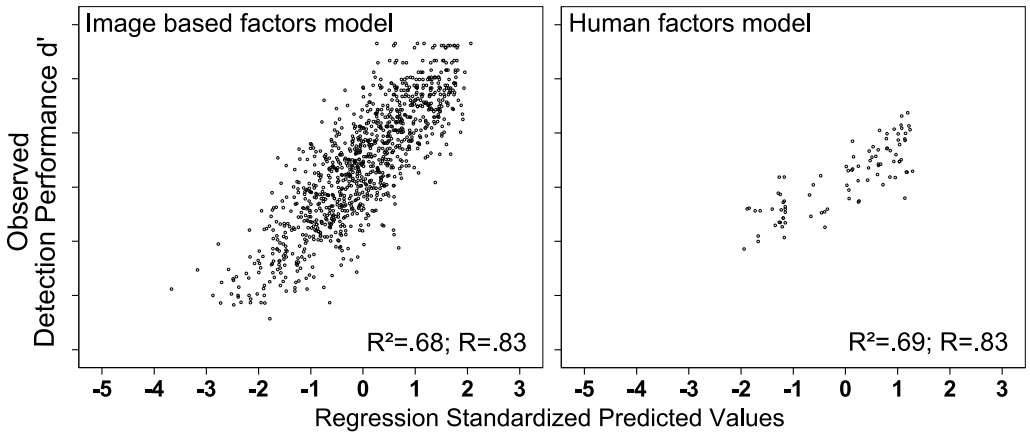

Figure 3.3: General regression models across all threat item categories. Scatterplots with standardized predicted values of the image based and human factors multiple linear regression models on the x-axis and observed detection performance d' on the y-axis.

between image based factors and threat detection performance with the relationship between human factors and threat detection performance separately for each threat category. Table 3.3 shows the most important statistical values for each of the reported models.

Image based factors

Figure 3.4 shows the four scatterplots illustrating the predictive power of the image based factors regression model separately for each threat category.

Human factors

Figure 3.5 shows the four scatterplots illustrating the predictive power of the human factors regression model separately for each threat category.

3.4.3 ANCOVA

Figure 3.6 shows a short overview of the ANCOVA results. The ANCOVA allows us to integrate human factors as covariates into a repeated measures ANOVA of image based factors (including threat category) and thus allows us to explore interaction effects and

Model Summaries (All Categories)				
		Predictors	Beta weights β	Significance p
Image Based Factors		FTI View Difficulty	$-.70$	.000
		logSuperposition	$-.127$	.000
		Opacity	$-.329$	.000
		Clutter	.198	.000
		Bag Size	.021	.288
		$R^2 = .68$, adjusted $R^2 = .68$, $F(5, 1018) = 441$, $p < .000$		
Human Factors		logTrainingHours	.93	.000
		Age	$-.26$	.000
		$R^2 = .69$, adjusted $R^2 = .68$, $F(2, 87) = 98$, $p < .000$		

Table 3.2: Tabular summary of the general multiple linear regression models for all threat item categories. Standardized beta weights and p-values.

dependencies among not just the image based factors, but also between the image based factors in combination with the human factors. On the left, Figure 3.6 illustrates the importance of the image based factors in terms of their effect size values η^2 (eta squared) and their interactions with the human factors. The main effects of each image based factor are reported together with their interaction effects with Training (log-transformed training hours) and Age, respectively. On the right, Figure 3.6 additionally illustrates the ten largest significant interactions in terms of η^2 values. For details on the data, their patterns and conclusions please consider Table 3.4 and the Discussions section at the end of this article.

Table 3.4 shows all η^2 values and the significance levels of the main effects and the interaction effects illustrated in Figure 3.6.

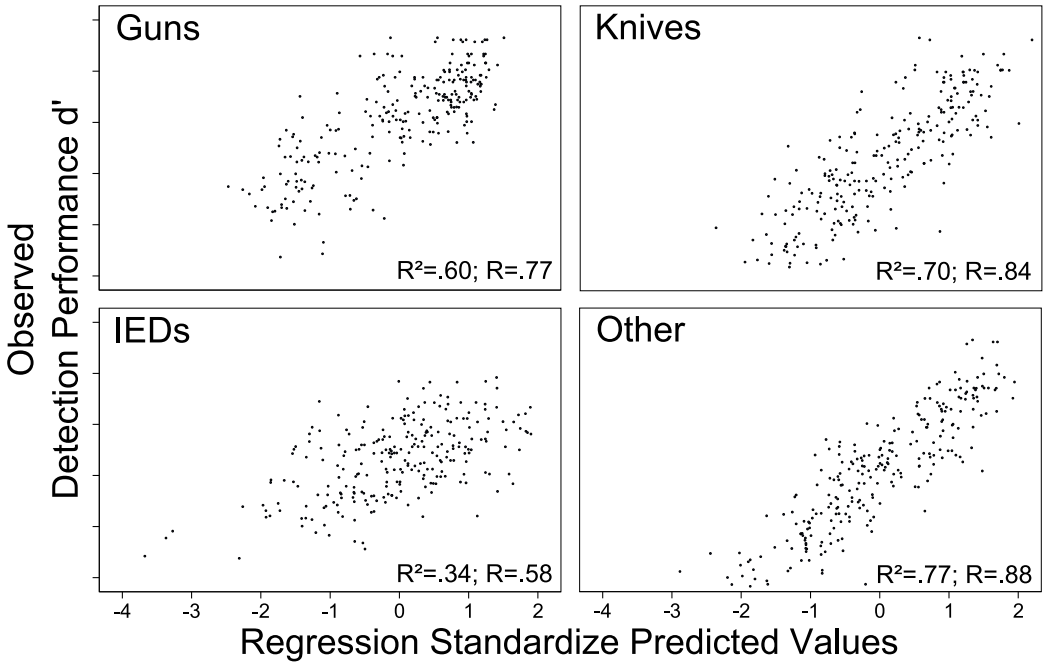

Figure 3.4: Separate image based factors regression models for each of the four threat item categories.

3.5 Discussions

For the discussion of our findings we retain the same presentation order as we did in the Methods and Results sections. Starting with the findings of the bivariate correlations we continue discussing the multiple linear regression models ending with the discussion of ANCOVA results.

3.5.1 Bivariate Correlations

The correlations between our factors and d' can be interpreted as the observed relationships between our predictors and d' observations. This gives us a first impression of how much explained variance we can expect from a single predictor. Figure 3.2 and Table 3.1 reveal that there are close relationships between d' and the three predictors FTI View Difficulty,

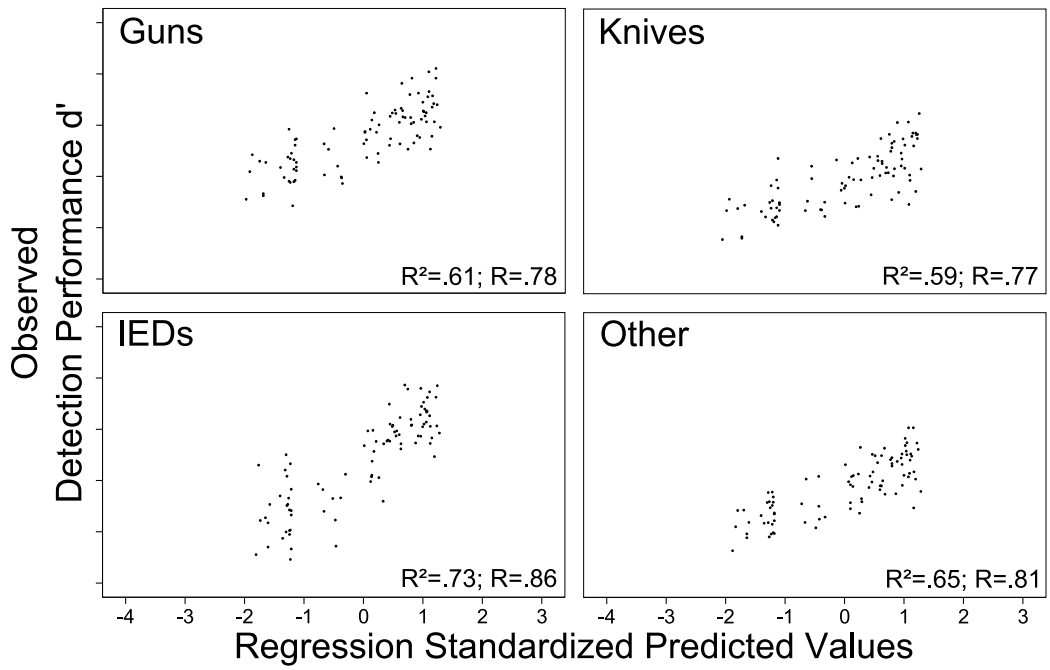

Figure 3.5: Seperate human factors regression models for each of the four threat item categories.

Superposition (log-transformed) and Training (log-transformed training hours). Also we can report a notable (partial) correlation between Age and d'. The remaining three predictors Opacity, Clutter and Bag Size show poor correlations. Clutter and Bag Size do not even reach the level of statistical significance of $p = .05$, Bag Size in knives being the only exception ($p < .01$). Table 3.1 shows both the bivariate- and partial correlations of the human factors with d' in order to allow direct comparisons. There are only very slight changes between bivariate and partial correlations with Training, but note that the signs of the correlations with Age all change from positive to negative when calculating partial correlation. The bivariate correlations reveal that the participants improve their detection performance with increasing Age. This finding contradicts earlier studies on visual search tasks that revealed a deterioration of performance with Age (Madden, Gottlob, & Allen, 1999). The partial correlations put this in perspective: Participants compensated age with more training, and indeed when controlling for Training there is a small negative correla-

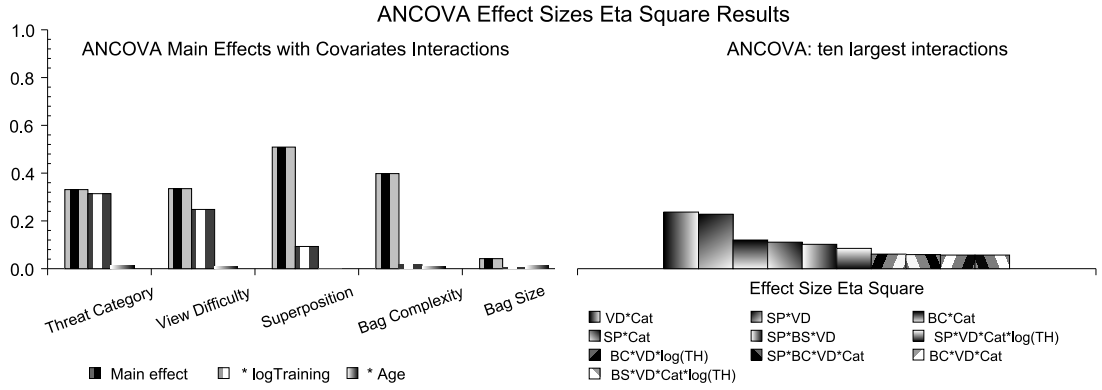

Figure 3.6: Summary of ANCOVA main effect and interaction effect sizes. All covariate interactions and the ten largest remaining interactions are reported.

tion between Age and detection performance.

A very interesting aspect of this analysis is the comparison of the correlations' patterns for the different threat categories. The data reveal that different factors are important for being able to identify FTIs belonging to different threat categories. Figure 3.2 illustrates impressively how View Difficulty has a similar amount of influence on detection performance of guns and knives, but a much smaller amount on IEDs and a notably larger one on Other. Interpreting the pattern of Other is very difficult because Other is just the rest category for all the threat objects that do not fit into any of the other three categories[1]. Thus the category Other includes as diverse objects as throwing stars (shuriken), tasers, hand grenades or gas tanks. Even though the IED stimulus set contains all sorts of hand made bombs, the stimuli are still comparatively homogenous making an interpretation of the image based factors much simpler than with Other: IEDs are generally made up of multiple essential parts (explosive material, fuse, cables, energy source, timer, etc.). Each of these has its own rotation (View Difficulty) and its own Superposition value. Therefore it is not too surprising that the effects of View Difficulty and Superposition are highly diminished. A very interesting complementary finding is that while for IEDs image based factors in general show the low-

[1]Threat categorization was inherited from the official ECAC threat categorization. ECAC refers to European Civil Aviation Conference

est impact on detection performance compared to the other threat categories, human factors - especially Training - show the strongest effects on d' with IEDs.

3.5.2 Multiple Linear Regression Models

As already anticipated in the Results section we are very happy to report the achieved explained variances of nearly 70% for both the image based factors regression model as well as for the human factors regression model. The fact that we are able to explain such a big portion of variance from two distinct sets of predictors independently makes us very confident to get a grip on the process of X-ray threat detection tasks. We are very confident that the image based factors together with human factors constitute the key aspects to cover for a better general understanding of the cognitive processes involved in this kind of visual search task. Nevertheless we still see some potential to further augment our model fits. This applies to the image based factors model as well as to the human factors model. Particularly with regards to the implementation of Clutter we see great potential to elicit larger predictive power - though to date we have not yet found a mathematical formula that rendered better results than the one currently in use. As for human factors: Until now we have merely investigated Training and Age. We see great potential in enlarging the set of human factors. Besides the undoubtedly important factor Training we expect visual abilities such as mental rotation, figure-ground segregation or visual search for highly specific patterns to be another important factor that should not be disregarded in a comprehensive model. Unfortunately for this study appropriate data were not available. As a standard human factor Gender should also be taken into account in future research.

Regarding the differences in terms of the explained variances and how they are made up of in terms of the beta weights among categories some remarkable notes must be made. The first pattern to catch one's eye is the different behavior of IEDs compared to the other threat categories. In the case of IEDs the image based factors' model regarding d' shows a

comparatively low amount of explained variance. For all other categories the image based factors have a very high predictive power. A closer look at the beta weights and correlations reveals that it is particularly the effect of FTI View Difficulty which lies far below what would be expected based on the results from the other threat categories. The complementary finding is that compared to the other categories IEDs show the largest amount of explained variance with the human factors model. The beta weights and correlations reveal that this can overwhelmingly be attributed to Training. We can conclude that IEDs depend largely on human factors, especially on Training. We assume that detection performance with IEDs depends on knowledge as opposed to visual abilities to a larger extent than is the case with the other categories. In general the human factor models show much lower differences in predictive power among threat categories than the image based factor models do. We assume that the explanation for the comparatively large variation of the overall predictive power of the image based models as well as the comparatively large amount of variation between the correlations of the individual image based factors with d' lies in detection performance being based on several distinct cognitive processes dealing with the different image based factors, with their relative importance varying between categories.

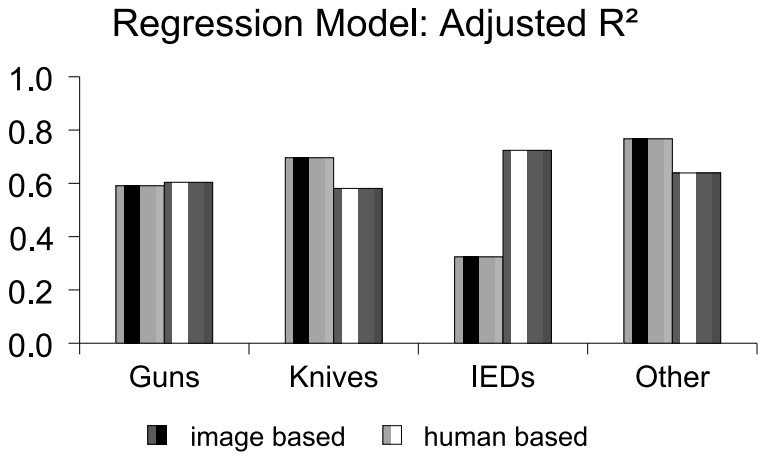

Figure 3.7: Comparison of adjusted R^2 values of image based factor and human factor linear regression models for each threat item category.

3.5.3 ANCOVA/ Interactions

There are some essential differences concerning data types. The factor View Difficulty as it is operationalized for the ANCOVA cannot be directly compared with the variable FTI View Difficulty used in the correlation analyses and the regression models. FTI View Difficulty is a proportionally scaled measure derived statistically from performance data whereas the ANCOVA factor View Difficulty is a dichotomous nominal variable differentiating between easy and difficult views only. As a matter of fact, all image based factors - except of course for the factor 'Threat Category' - are dichotomous in the ANCOVA (refer to the paragraph on stimuli in the Methods section). Bag Complexity replaces the compound of Opacity and Clutter (refer to the Stimuli section).

The main effects of the ANCOVA on all categories give a similar picture as did the correlations. In analogy to what we found in the correlation analyses, effects of Superposition are larger than effects of Bag Complexity and Bag Size. Furthermore the results show that there are considerable interactions with Training regarding the different threat categories. For example, as discussed above the effect of Training on detection performance is clearly larger in IEDs than in knives.

View Difficulty and Superposition also show interaction effects with Training. The improvement of the detection performance caused by Training is clearly larger regarding the difficult views compared to the easy ones. The interaction between Superposition and Training on the other hand is fairly small. This could indicate that dealing with superposition is difficult to improve with Training. For Bag Complexity and Bag Size interactions with Training are not significant. No evidence could be provided for interaction effects of any of our image based factors (including threat category) with Age.

In the following itemization we give a short overview of the four largest reported first order interactions and plausible interpretations with examples.

	Model Summaries (Per Category)							
	Predictors	Guns		Knives		IEDs		Other	
		β	p	β	p	β	p	β	p
Image Based Factors	FTI View Difficulty	$-.67$	.000	$-.67$	.000	$-.29$	.000	$-.79$	.000
	logSuperposition	$-.12$	.025	$-.18$	.000	$-.25$	.000	$-.09$	.061
	Opacity	$-.27$	.000	$-.36$	.000	$-.51$	.000	$-.31$	.000
	Clutter	.18	.005	.17	.003	.35	.000	.20	.000
	Bag Size	.04	.375	$-.05$	.256	.09	.105	.05	.187
	R^2	.60		.70		.34		.77	
	$adjusted R^2$	.59		.70		.32		.77	
	$F(5, 250)$	75		117		25		169	
	p	.000		.000		.000		.000	
Human Factors	logTraining	.88	.000	.88	.000	.94	.000	.90	.000
	Age	$-.26$	.001	$-.29$	.000	$-.20$	.003	$-.22$	.004
	R^2	.61		.59		.73		.65	
	$adjusted R^2$	.60		.58		.72		.64	
	$F(2, 87)$	69		63		118		80	
	p	.000		.000		.000		.000	

Table 3.3: Tabular summary of separate multiple linear regression models for each of the four threat item categories. Standardized beta weights and p-values.

ANCOVA effect sizes η^2					
	Category	View Difficulty	Superposition	Bag Complexity	Bag Size
Main effects	.33***	.34***	.51***	.40***	.04
* logTraining	.31***	.25***	.09**	.02	.00
* Age	.01	.01	.00	.01	.01

Table 3.4: Summary of ANCOVA main effects and covariate interactions.

VD * Cat: reflects the fact that effects of View Difficulty differ between Threat Categories e.g.: compare correlations of View Difficulty and d' between IEDs and Other

VD * SP: difficult views are more affected by high Superpositions that easy ones and vice versa

BC * Cat: reflects the fact that effects of Bag Complexity differ between Threat Categories

SP * Cat: reflects the fact that effects of Superposition differ between Threat Categories

Chapter 4

Model Consolidation II - Completing the Circle

4.1 A Statistical Approach for Image Difficulty Estimation in X-Ray Screening Using Image Measurements

The relevance of aviation security has increased dramatically at the beginning of this century. One of the most important tasks is the visual inspection of passenger bags using X-ray machines. In this study, we investigated the role of image based factors on human detection of prohibited items in X-ray images. Schwaninger et al. (2004); Schwaninger, Hardmeier, and Hofer (2005) have identified three image based factors: View Difficulty, Superposition and Bag Complexity. This article consists of 5 experiments which lead to the development of a statistical model that is able to predict image difficulty based on the mentioned image based factors. Experiment 1 is a replication of earlier findings confirming the relevance of image based factors as defined by Schwaninger, Hardmeier, and Hofer (2005) on X-ray detection performance. In Experiment 2, we found significant correlations between human

ratings of image based factors and human detection performance for most of the image based factors. In Experiment 3, we introduced our image measurements and found mostly significant correlations between them and human detection performance. Moreover, significant correlations were found between our image measurements and corresponding human ratings, indicating high perceptual plausibility. In Experiment 4 we contrasted the image based factors measurements with the human ratings from Experiment 2 in terms of their predictive power in multiple linear regression models. Experiment 5 is another replication of the image measurements part in Experiment 4, whereby the data are based on a much more extensive test as well as on a larger sample size of professional screeners. Applications of a computational model for threat image projection systems and for adaptive computer-based training are discussed.

4.2 Introduction

The relevance of aviation security has increased dramatically in recent years and there has been substantial progress regarding screening technology, especially in the field of automatic explosive detection systems (Ying et al., 2006). However, the last decision is always made by a human operator and investigating human factors as essential determinants of security screening performance has become an important research topic. First contributions in the field of X-ray image inspection were based on research in medical image interpretation (Gale et al., 2000). Krupinski et al. (2003) were able to identify important factors that influence pulmonary nodule detection. Experimental psychology studies (Ghylin et al., 2006) and eye movement research (McCarley et al., 2004; Liu et al., 2006) have been useful to better understand visual search and perceptual learning in X-ray image interpretation. A series of studies conducted in recent years has provided converging evidence for the importance of scientifically based selection, training and testing methods to achieve and maintain high levels of performance in X-ray image interpretation (Schwaninger, 2005b,

2006b).

The aim of this study is to develop and evaluate a statistical model for image difficulty estimation in X-ray screening using image measurements. Schwaninger, Hardmeier, and Hofer (2005) could show that there are three major image based factors which affect detection performance: View difficulty depending on the rotation of an object, Superposition by other objects in the bag, and Bag Complexity, which comprises Clutter, the bag's background texture unsteadiness, and Opacity, the relative size of dark areas in the bag. Figure 4.1 illustrates the three image based factors as proposed by Schwaninger, Hardmeier, and Hofer (2005).

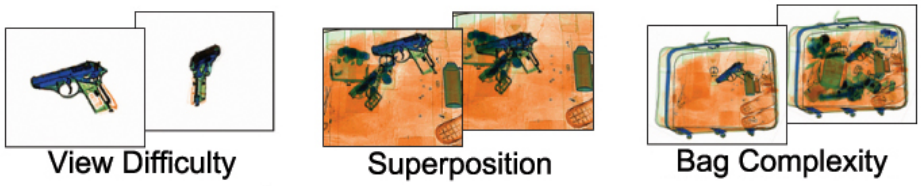

Figure 4.1: Illustration of the three major image based factors suggested by Schwaninger et al. (2004); Schwaninger, Hardmeier, and Hofer (2005).

A model for image difficulty estimation using automated image measurements and human performance statistics can be very useful for threat image projection (TIP) data analysis and individually adaptive computer based training (CBT) systems. TIP is a software function of state-of-the art X-ray machines which allows automated insertion of fictional threat items (FTIs) into X-ray images of real passenger bags. TIP systems are operational in several countries and used to enhance motivation and attention of screeners on the job. Since the TIP to bag ratio is relatively low (i.e. the number of projections per passenger bags) and the resulting TIP images (X-ray image of real passenger bag plus FTI) vary substantially with regard to image based factors, it is difficult to obtain reliable individual performance measurements. With a reliable statistical model for image difficulty estimation using image measurements, corrected individual performance scores could be calculated, which would allow more reliable individual performance assessments. A second application is

adaptive CBT. For example, in the computer based training program X-Ray Tutor, the individually adaptive algorithms start displaying easy views of threat items shown in bags of low complexity with little Superposition by other objects. Once a threat item is recognized by a screener, the View Difficulty is increased and it is shown in more complex bags with more Superposition (for details on X-Ray Tutor see Schwaninger (2004b)). There are large differences between individuals regarding their ability to cope with image-based factors Schwaninger, Hardmeier, and Hofer (2005). Therefore, a good model for image difficulty estimation using automated image measurements of image-based factors could be very useful for enhancing such individually adaptive training algorithms.

The study is sectioned into five experiments. The first experiment is a replication of earlier findings (Schwaninger, Hardmeier, & Hofer, 2005) to confirm the relevance and relative independence of image based factors in predicting human performance. The second experiment aims to estimate the subjective perceptual plausibility of the underlying image based factors by correlating them with the average d'. Threat images were rated for View Difficulty, Superposition, Clutter, Opacity and general difficulty. Images of harmless bags were rated for Clutter, Opacity, and general difficulty only. The correlation between these ratings and human detection performance reflects the relative importance of each image based factor. We then developed statistical formulae and automated image measurements for the above mentioned image based factors. Experiment 3 was designed to estimate the perceptual plausibility of these computer generated estimates. We correlated the computer-based estimates with the corresponding human ratings to determine whether our computer-based algorithms correspond with human perception. In Experiment 4 we compared a model using computer-based estimates to a model based on human ratings of the image based factors. Experiments 1 to 4 are concerted and are based on the same subjects, images and tests. Finally, Experiment 5 provides a fairly extended replication of the calculations model in Experiment 4. The mentioned extension in Experiment 5 refers to the inclusion of Bag Size as an additional image based factor. Further, the data used in Experiment 5 stem from

completely different subjects (in terms of sample size) as well as in terms of the test used and its image stimuli. The larger data base underlying Experiment 5 results in much more stable and clearer results than the ones achieved in Experiment 4.

4.3 Experiment 1

4.3.1 Method

Experiment 1 is a replication of the study by Schwaninger, Hardmeier, and Hofer (2005), who identified image based factors for threat item detection in X-ray image screening. In all of the following experiments the dependent variable used is the signal detection measure d' (Green & Swets, 1966). Unlike the popular detection performance hit rate, d' takes into account also the false alarm rate: $d' = z(H) - z(FA)$ whereas $z(H)$ refers to the z-transformed hit rate and $z(FA)$ to the z-transformed false alarm rate. As a consequence d' can be considered as criterion independent, i.e. d' values are independent of whether a screener behaves very risky and efficiently or whether the screener is very anxious.

Participants

Nineteen highly experienced aviation security experts from a large European airport participated in this experiment (10 females). Due to security reasons and data protection no absolute d'-values, hit rates and false alarm rates are reported.

Materials

The X-Ray Object Recognition Test (X-Ray ORT) was used to measure detection performance. This test has been designed to analyze the influence of image based effects View

Difficulty, Superposition and Bag Complexity on human detection performance when visually inspecting X-ray images of passenger bags. Inspired by signal detection theory (Green & Swets, 1966), the X-Ray ORT consists of two sets of 128 X-ray images. One set contains harmless bags without a threat item (N-trials, for noise). The other set contains the same bags, each of them containing a threat object (SN-trials, for signal-plus-noise). Only guns and knives of typical familiar shapes are used. This is important because the X-Ray ORT is designed to measure cognitive visual abilities to cope with effects of View Difficulty, Superposition, and Bag Complexity independent of specific visual knowledge about threat objects. Due to the same reason a greyscale version of the X-Ray ORT was used for this study. The X-Ray ORT consists of 256 items (X-ray images) given by the following test design: 16 threat item exemplars (8 guns, 8 knives) x 2 View Difficulty levels x 2 Bag Complexity levels x 2 Superposition levels x 2 trial types (SN and N-trials). The construction of the items in all image based factor combinations as shown above was lead by visual plausibility criteria. After choosing two sets of X-ray images of harmless bags with different parameter values in Bag Complexity, the sixteen fictional threat items were projected into the bags in two different view difficulties at two locations with different Superposition each. The term fictional threat items (FTIs) is commonly used in connection with TIP systems as discussed in the introduction. For further details on the X-Ray ORT see Hardmeier et al. (2005); Schwaninger, Hardmeier, and Hofer (2005). Stimuli were displayed on 17" TFT screens at a distance of about 100cm, so that the X-ray images subtended approximately 10-12 degrees of visual angle. The computer program measured outcome (hit, miss, false alarm, correct rejection) and the response times from image onset to final decision button press.

Procedure

X-ray images of passenger bags were shown for a maximum display duration of 4 seconds. Note that at airport security controls the human operators (screeners) usually have only 3-6

seconds to inspect a passenger bag. The participant's task was to decide whether the image is OK (i.e. the bag contains no threat item) or NOT OK (i.e. it contains a threat item) by clicking one of the corresponding buttons on the screen (see Figure 4.2). In addition, participants had to judge their confidence using a slider control (from UNSURE to SURE). The confidence ratings are not used in this study. No feedback was given regarding the correctness of the responses. Participants could initiate the next trial by pressing the space bar.

Figure 4.2: Screenshot of an X-Ray ORT trial showing an X-ray image of a passenger bag containing a gun. Response buttons and slider control are aligned at the bottom of the screen.

Several practice trials were presented to make sure that the task was understood properly before the test started. Immediately prior to the actual test, all guns and knives were presented on the screen for 10 seconds each threat category. This was done to minimize any effects of threat item knowledge. Half of the items were shown in easy view and the other half in difficult view.

4.3.2 Results

Figure 4.3 displays the mean d'-values and standard deviations broken up by main effects of View Difficulty, Superposition, and Bag Complexity for guns and knives separately. Data was first averaged across images for each participant and then across participants to

calculate mean hit rate. Separate analyses of hit rates and false alarm rates are considered to be published later.

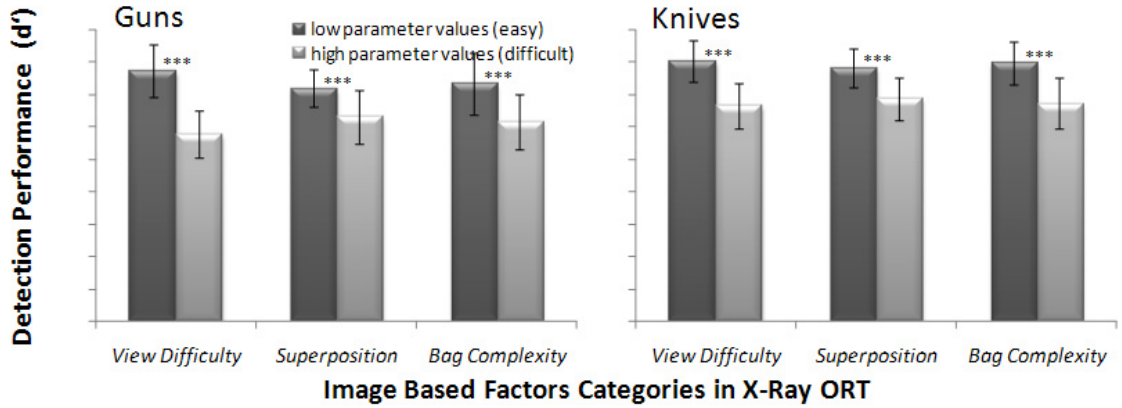

Figure 4.3: Results of Experiment 1. Mean d' detection performance of guns and knives, broken up by main effects of View Difficulty, Superposition, and Bag Complexity. Data was first averaged across images for each participant and then across participants to calculate mean d'. Error bars represent the standard deviation across participants. Due to data protection reasons, absolute d'-values are not published.

Our hypothesis whereby the image based factors have great influence on detection performance was tested using repeated-measures ANOVA. Table 4.1 shows the main effects of the image based factors for guns and knives separately. Effect sizes are given as η^2 values. All image based factors effects are highly significant.

4.3.3 Discussion

We were able to replicate the results from Schwaninger, Hardmeier, and Hofer (2005) and Hardmeier et al. (2005) very well. As mentioned earlier, d' equals $z(H) - z(FA)$ whereas H refers to hit rate and FA to false alarm rate (Green & Swets, 1966). Schwaninger, Michel, and Bolfing (2007) showed in their APGV paper that effects of Bag Complexity did not get significant when taking the hit rate as dependent variable alone. Only View Difficulty and Superposition had significant effects on the hit rate. Effects of Bag Complexity

Table 4.1: Repeated Measures ANOVA Main Effects

Guns:			
View Difficulty:	$\eta^2 = .86$	$F(1, 18) = 110.41$	$p < .001$
Superposition:	$\eta^2 = .64$	$F(1, 18) = 32.44$	$p < .001$
Bag Complexity:	$\eta^2 = .45$	$F(1, 18) = 14.83$	$p < .001$
Knives:			
View Difficulty:	$\eta^2 = .77$	$F(1, 18) = 60.89$	$p < .001$
Superposition:	$\eta^2 = .70$	$F(1, 18) = 42.30$	$p < .001$
Bag Complexity:	$\eta^2 = .63$	$F(1, 18) = 30.22$	$p < .001$

are more likely to be found on false alarm rate. In X-ray screening tests, the false alarm rate is based on the number of times a participant judges a bag to be NOT OK even though there is no threat item in it. Since View Difficulty and Superposition only exist with bags containing a threat item these findings are not too surprising.

4.4 Experiment 2

Experiment 2 was designed to investigate the perceptual plausibility of our image measurements introduced in Experiment 3.

4.4.1 Method

Eleven out of the nineteen experts who had conducted Experiment 1 took part in Experiment 2 showing a modified experimental setup. The participant's task was to rate the difficulties of the X-Ray ORT images regarding View Difficulty and Superposition of the threat images. In addition, Clutter, Opacity and general item difficulty had to be rated for

threat and non-threat images. As you noticed, the image based factor Bag Complexity was replaced by two new image based factors Clutter and Opacity representing two different aspects of Bag Complexity. This splitting of Bag Complexity was necessary to address the two aspects separately in the development of the automatic image measurement algorithms described in Experiment 3. The ratings were given by mouse clicks on a 50-point scale (0 = very low to 50 = very high). No initial position was set. Figure 4.4 shows a screenshot.

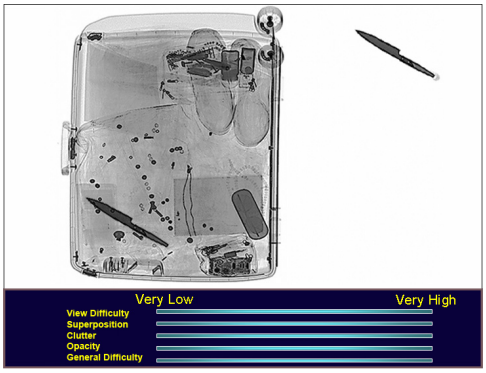

Figure 4.4: Screenshot of a typical trial of Experiment 2. All subjects were asked to judge the image based factors subjectively, whereby Bag Complexity was separated in Clutter and Opacity. Additionally, participants were asked to judge the general item difficulty as well (not analyzed in this study). Threat items were displayed next to the bag. For non-threat items, the slider controls for View Difficulty and Superposition were discarded.

4.4.2 Results

In order to estimate the relative importance of image based factors (Schwaninger, Hardmeier, & Hofer, 2005) on human detection performance, we correlated ratings for View Difficulty, Superposition, Clutter and Opacity (Experiment 2) with the hit rate data obtained in Experiment 1. Data analysis was conducted separately for guns and knives.

Figure 4.5 shows the averaged ratings across all participants and across all threat items. The ordinate depicts the rating scores on the 50-point scale (see Figure 4.4). The dark and bright gray bars in each image based factors category represent the low and high parameter

values according to the arrangement of the X-Ray ORT test design. Inter-rater consistency was quite high with an average correlation (Fisher-corrected) between subjects of $r = .61$ for View Difficulty, $r = .60$ for Superposition, $r = .62$ for Clutter and $r = .33$ for Opacity.

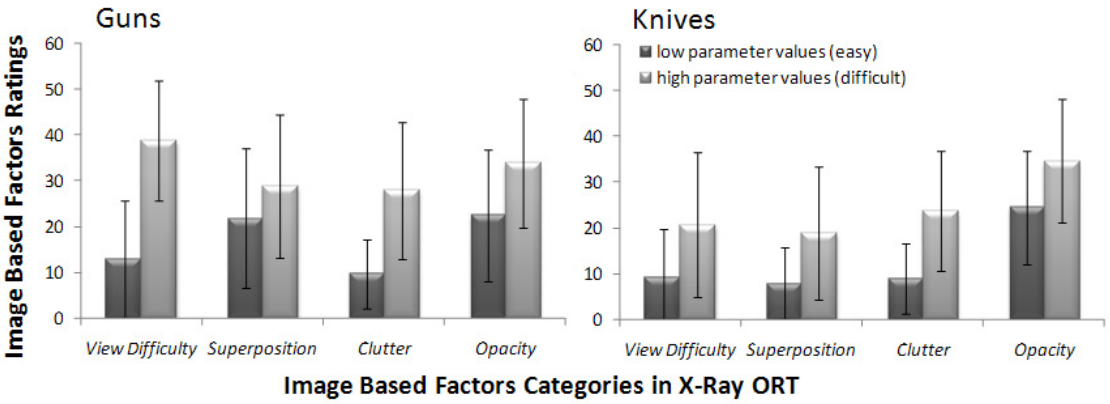

Figure 4.5: Descriptive results from Experiment 2 for guns and knives separately. The image based factor Bag Complexity from the X-Ray ORT is split into the sub-factors Clutter and Opacity according to the rating experiment design shown in Figure 4.4.

Table 4.2 shows the correlations between the averaged (across subjects) human ratings of the image based factors and d'-values per image.

Concerning the mathematical signs, note that d' points in the opposite direction of threat detection difficulty. The more difficult a threat item is to be detected the lower d'.

4.4.3 Discussion

Most subjective human ratings show correlations with d' from Experiment 1 with quite low significance levels even though many correlation did get significant at the 5% level. Thus, the results from Experiment 1 and Experiment 2 showed that image based factors affect objective X-ray image difficulty (d') and that the image-based factors can be rated quite consistently. For the development of image measurements, it was necessary to split up the factor Bag Complexity into Clutter and Opacity. However, this seems to be problematic,

Table 4.2: Correlation between Human Ratings and d'

Guns:		
View Difficulty:	$r(64) = -.46$	$p < .001$
Superposition:	$r(64) = -.66$	$p < .001$
Opacity:	$r(64) = -.21$	$p = .095$
Clutter:	$r(64) = -.25$	$p < .05$
Knives:		
View Difficulty:	$r(64) = -.24$	$p = .06$
Superposition:	$r(64) = -.61$	$p < .001$
Opacity:	$r(64) = -.32$	$p < .05$
Clutter:	$r(64) = -.17$	$p = .173$

because for subjective ratings they seem to be highly interdependent. The ratings of Clutter and Opacity are highly correlated: $r(64) = -.93, p < .001$ for both guns and knives. We return to this issue in section 4.3.

4.5 Experiment 3

The aim of Experiment 3 was to develop computer-based algorithms to automatically estimate the image based factors View Difficulty, Superposition, Clutter, and Opacity. Unlike in earlier studies dealing with the same image based factors, we decided to rename the factor transparency into Opacity to take into account the mathematical sign of the image measurement formula as described below as well as to fit increasing image based factor values to increasing detection difficulties. The perceptual plausibility of these computer-based algorithms was examined by correlating them with the human ratings obtained in Experiment 2.

4.5.1 Statistical Estimates and Image Measurements for Image Based Factors

All image measurements developed for this purpose are based on theoretical considerations. Different algorithm parameters were optimized by maximizing the correlations between the image-based factors estimates and detection performance measures derived from earlier X-Ray ORT findings. In the following all four image based factors are described separately regarding their concepts, statistics and image measurement calculations:

FTI View Difficulty

Even with the aid of 3D volumetric models, it is not (yet) possible to satisfyingly determine the degree of a 3-dimensional rotation (View Difficulty) of a physical threat item automatically from its 2-dimensional X-ray image (Mahfouz et al., 2005). Additional difficulties regarding image segmentation arise from the very heterogeneous backgrounds of X-ray images, compare (Sluser & Paranjape, 1999). Therefore, this image based factor is not (yet) being calculated by image processing, but post hoc from X-Ray ORT detection performance data obtained in Experiment 1.

$$\text{FtiVD}_{OVj} = \frac{\sum_{i=1,j\neq i}^{4} (4.65 - d'_{OVi})}{3} \tag{4.1}$$

Equation 4.1 shows the calculation of the image based factor FTI View Difficulty, whereas i is the summation index ranging from 1 to 4 (2 bag complexities x 2 Superpositions), j denotes the index number of the X-ray image in question (one threat object (O) in one of the two views (V)), d'_{OVj} is its average d' across all participants and '4' is the number of the bags each FTI was projected into. In order to avoid a circular argument in the statistical model (multiple linear regression, see Experiment 4) by partial inclusion of the

criterion variable into a predictor, d' of the one item in question is excluded from this estimate. The constant term 4.65 in the equation represents the maximum possible d'-value and was introduced in order for FTI View Difficulty to point in the same direction as image difficulty does. It is important to understand that this concept of FTI View difficulty is not just reflecting the degree of rotation of an object. View difficulty as it is conceptualized here reflects innate FTI View Difficulty attributes unique to each FTI view separately.

Superposition

This image based factor refers to how much the pixel intensities at the location of the FTI in the threat bag image differ from the pixel intensities at the same location in the same bag without the FTI. Equation 4.2 shows the image measurement formula for Superposition. $I_{SN}(x, y)$ denotes the pixel intensities of a threat image and $I_N(x, y)$ denotes the pixel intensities of the corresponding harmless bag. The subtraction of the square root term from a constant is used in order for Superposition to point in the same direction as image difficulty does. In this analysis the constant term C was set to 0.

$$SP = C - \sqrt{\sum_{x,y} \left(I_{SN}(x, y) - I_N(x, y)\right)^2} \tag{4.2}$$

It should be noted that this mathematical definition of Superposition is dependent on the size of the threat item in the bag. For further development of the computational model it is conceivable to split up Superposition and the size of the threat item into two separate image based factors. Measurement of Superposition would require having both the bag with the FTI and without it. For both applications mentioned in the introduction, this is possible with current TIP and CBT technology. In TIP, the FTI, its location, the bag with and without the FTI are recorded. In several CBT systems, the same information is recorded and stored, too.

Clutter

This image based factor is designed to express bag item properties like its textural unsteadiness, disarrangement, chaos or just Clutter. In terms of the bag images presented, this factor is closely related to the amount of items in the bag as well as to their structures in terms of complexity and fineness. The method used in this study is based on the assumption, that such texture unsteadiness can be described mathematically in terms of the amount of high frequency regions.

$$CL = \sum_{x,y} I_{hp}(x, y) \tag{4.3}$$

$$\begin{aligned} \text{where} \quad I_{hp}(x, y) &= I_N * \mathcal{F}^{-1}(hp(f_x, f_y)) \\ &= \mathcal{F}^{-1}(\mathcal{F}(I_N \cdot hp(f_x, f_y)) \end{aligned}$$

Equation 4.3 shows the image measurement formula for Clutter. It represents a convolution of the empty bag image (N for noise) with the convolution kernel derived from a high-pass filter in the Fourier space. I_N denotes the pixel intensities of the harmless bag image. $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. $hp(f_x, f_y)$ represents a high-pass filter in the Fourier space (see Appendix).

Opacity

The image based factor Opacity reflects the extent to which X-rays are unable to penetrate objects in a bag. This depends on the specific material density of these objects. These attributes are represented in X-ray images as different degrees of luminosity. Heavy metallic materials such as lead are known to be very hard to be penetrated by X-rays and therefore appear as dark areas on the X-ray images.

$$OP = \frac{\sum_{x,y}(I_N(x,y) < 64)}{\sum_{x,y}(I_N(x,y) < 252)} \tag{4.4}$$

Equation 4.4 shows the image measurement formula for Opacity. $I_N(x, y)$ denotes the pixel intensities of the harmless bag. 64 is the pixel intensity threshold beneath which the pixels are counted. The implementation of the image measurement for the image based factor transparency is simply achieved by counting the number of pixels being darker than a certain threshold (< 64) relative to the bag's overall size (< 252, non-white pixels).

4.5.2 Method

To get a feeling of the relations between our image measurement estimates, the human ratings and detection performance d' the correlations between the image measurement estimates and d' are presented in the following results section. To examine perceptual plausibility of the computer based measurements, we correlated them with the human ratings from Experiment 2.

4.5.3 Results

Correlation between Image Measurements and Detection Performance d'

Pearson's product-moment correlations between the calculated measurements and d' from Experiment 1 were applied for each image based factor dimension separately. Table 4.3 shows the respective r-values and the corresponding significance levels.

Table 4.3: Correlations between the Image Measurements and d'

Guns:		
View Difficulty:	$r(64) = -.59$	$p < .001$
Superposition:	$r(64) = -.38$	$p < .01$
Opacity:	$r(64) = -.33$	$p < .01$
Clutter:	$r(64) = -.08$	$p = .533$
Knives:		
View Difficulty:	$r(64) = -.39$	$p < .01$
Superposition:	$r(64) = -.39$	$p < .01$
Opacity:	$r(64) = -.31$	$p < .05$
Clutter:	$r(64) = -.29$	$p < .05$

Correlation between Image Measurements and Human Ratings

Pearson's product-moment correlations between the calculated measurements and the corresponding human ratings' mean values were applied for each image based factor dimension separately. Table 4.4 shows the bivariate correlations and their significance levels.

4.5.4 Discussion

Except for Clutter in guns all correlations between automated image measurements and detection performance are significant. The correlation can be referred to as the direct effect sizes of each of the image based factors measurement d'. For the over-all predictive power of the measurements please refer to Experiment 4.

Except for Clutter almost all correlations between automated measurements and ratings are significant. In the discussion of Experiment 2 the high inter-correlations between the human ratings of the image based factors Clutter and Opacity was mentioned ($r(64) =$

Table 4.4: Correlations between the Image Measurements and the Human Ratings

Guns:		
View Difficulty:	$r(64) = .54$	$p < .001$
Superposition:	$r(64) = .53$	$p < .001$
Opacity:	$r(64) = -.69$	$p < .001$
Clutter:	$r(64) = .15$	$p = .25$
Knives:		
View Difficulty:	$r(64) = .20$	$p = .113$
Superposition:	$r(64) = .39$	$p < .01$
Opacity:	$r(64) = -.64$	$p < .001$
Clutter:	$r(64) = .08$	$p = .516$

$-.93, p < .001$ for both guns and knives). Consistent with this result, there were also fairly high inter-correlations between the corresponding calculated estimates of the image based factors Clutter and Opacity ($r(64) = .42, p < .001$ for guns and $r(64) = .36, p < .01$ for knives). Except for Clutter, we can conclude that our algorithms for automated estimation of image based factors are perceptually plausible because they correlate significantly with the ratings of novices.

4.6 Experiment 4

Experiment 4 was designed to evaluate the predictive power of a statistical model based on automated estimation of image based factors. To this end, we now compare the results of multiple linear regression analysis using the automated estimates of image based factors as predictors with the results of multiple linear regression analysis using the human ratings of image based factors as predictors.

4.6.1 Method

Multiple Linear Regression Analysis

The predictors of the multiple linear regression model are our image based factors introduced in Experiment 3. d' per image averaged across subjects (Experiment 1) is the dependent variable. We compared the two statistical models in terms of their goodness-of-fit measures, their regression coefficient's significances and the percentage of variance in the dependent variable d' the model is able to explain by its predictors.

4.6.2 Results

Figure 4.6 shows the scatter plots with regression standardized predicted values on the abscissa and the actually measured d' from Experiment 1 on the ordinate.

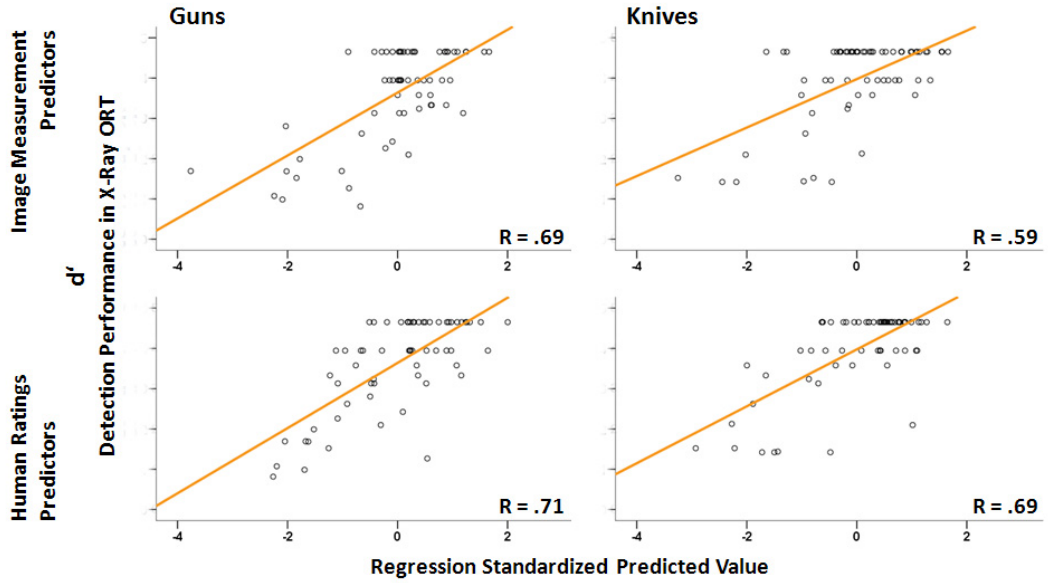

Figure 4.6: The four scatter plots from the models predicting d' on the basis of all disposable image based factors as predictors. Guns and knives are displayed separately. The models based on the calculated predictors derived from image measurements are displayed upon the models based on rated image based factors.

Table 4.5 shows the most important statistical values of the four multiple linear regression analyses arranged like in Figure 4.6. The single tables show the four predictors in the rows. The first column gives the variable names of the image based factors. Standardized beta weights are displayed in the second column and column three shows the p-value statistics indicating the significance of the single regression coefficients in the model. The last line shows the goodness-of-fit statistics of the model as a whole. R^2 tells us to which extent the model is able to predict the variance in d'. Because R^2 increases with the number of predictors independently of their predictive power, $R^2(adj)$ taking into account the number of predictors is given, too. Finally the statistical indices F-value and the significance level of the model as a whole (p-value) are given.

All statistical models are highly significant in the overall goodness-of-fit verification statistics, both for guns and knives. The R^2-values, the extent to which a model is able to explain the variance in the dependent variable by its predictors, are very high compared to values usually obtained when predicting human performance. The model based on our image measurements achieves an R^2 of .47 ($R^2(adj)$=.43) with guns and an R^2 of .35 ($R^2(adj)$=.30) with knives. The ratings model is even better with an R^2 of .51 ($R^2(adj)$=.47) with guns and an R^2 of .48 ($R^2(adj)$=.44) with knives.

4.6.3 Discussion

The different statistical models in Experiment 4 show that the image based factors suggested by Schwaninger, Hardmeier, and Hofer (2005) are quite powerful predictors of human detection performance. Such a model could therefore provide a basis for the enhancement of individually adaptive computer-based testing and training systems in which the estimation of X-ray image difficulty is an essential component. In addition, the image measurements developed in this study can be very useful for analyzing more reliable individual TIP performance scores by taking into account image difficulty as explained in the

Model Summaries (Per Category)					
		Guns		Knives	
		β -weights	p-values	β -weights	p-values
Image Measurements	View Difficulty	-.54	.000	-.29	.015
	Superposition	-.15	.117	-.31	.010
	Opacity	-.30	.008	-.19	.126
	Clutter	.03	.805	-.19	..126
	R^2	.47		.35	
	$R^2(adj)$	.43		.30	
	$F(4, 59)$	13		8	
	p	.000		.000	
Human Ratings	View Difficulty	-.27	.009	.05	.647
	Superposition	-.54	.000	-.56	.000
	Opacity	-.04	.883	.944	.001
	Clutter	-.11	.672	.86	.002
	R^2	.51		.48	
	$R^2(adj)$	.47		.44	
	$F(4, 59)$	10		14	
	p	.000		.000	

Table 4.5: Statistical analysis tables of the models with the most important statistical values of the multiple linear regression analyses. Each of the four tables shows the statistical values of the verification of each regression coefficient separately in the rows. Additionally the model's overall goodness-of-fit verification values are given at the bottom row of each model. In both statistical models, the dependent variable is the d' from Experiment 1.

introduction.

It is interesting to discuss the differences in the beta-weights between guns and knives in the image processing model. For guns View Difficulty is weighted almost double compared to Superposition. For knives the contrary pattern was observed. We are currently conducting additional analyses to find out whether this effect is related to differential changes by 3D rotation. The reason why Superposition is weighted much stronger in knives than in guns is probably due to the Superposition formula which also reflects the size of the threat items. In the X-Ray ORT knives differ much more in size than guns. Thus, the regression coefficient patterns reflect actual characteristics of the FTI categories. The scatter plots (Figure 4.6) reveal that there is a quite large ceiling effect. Therefore, it might be of value to use non-linear regression for modeling d' in the future. Apart from that, this study can be viewed as the basis for further statistical models for the prediction of individual screener responses to single X-ray images using binary logistic regression. In addition, together with the development of enhanced and additional image based predictors we intend to develop parallel statistical models to predict hit rates as well as false alarm rates.

4.7 Experiment 5

In Experiment 5 we sought to replicate and refine the results of Experiment 4 by increasing the number of test items and the subject sample size. A fifth image based factor, Bag Size, was added to the other four, FTI View Difficulty, Superposition, Opacity and Clutter.

4.7.1 Method

For Experiment 5 we did not use the ORT. Instead a new test, based on the X-Ray CAT (X-Ray Competency Assessment Test) by Koller and Schwaninger (2006), was designed.

We shall refer to this test as 'X-Ray Bag Size Test', because it is the first test we use which also measures Bag Size effects. In comparison to the X-Ray ORT, the X-Ray Bag Size Test shows color images and has been expanded to include two additional categories of prohibited items next to guns and knives: 'IEDs' (Improvised Explosive Devices) and 'other'. The Category other serves as a container for objects which do not fit the first three categories, such as pepper sprays or tasers. With regards to possible statistical analyses the X-Ray Bag Size Test has an analogous balancing of its image based factors FTI Category (guns, knives, IEDs or other), View Difficulty, Superposition, Bag Complexity and Bag Size as does the X-Ray ORT. To make the item sample even larger and allow a more precise analysis of the data, the X-Ray Bag Size Test contains 16 threat item exemplars per Category instead of 8. Among other things, the larger item sample would allow us to see, whether the relationships between the image based factors and detection performance d' were indeed linear (as is implicitly assumed by computing multiple linear regression models). Should this not be the case appropriate prior data-transformations could be applied prior. In total the test consisted of 2048 trials.

Participants

Trained as well as newly employed untrained screeners at a European airport employing X-Ray Tutor. The data of 63 screeners is used in our analysis (Age M = 33.9, SD = 14.1).

Materials

In contrast to the X-Ray ORT the X-Ray Bag Size Test is not designed to measure cognitive visual abilities independent of visual knowledge. Rather it is designed to provide a measures of detection performance with high ecological validity. The focus of the X-Ray Bag Size Test's design lay on providing a tool for reliable measurement of the different image based factors - including their interactions. The test consists of two sets of 1024 X-

ray images. One set contains harmless bags without a threat item (N-trials, for noise). The other set contains the same bags, each of them with a threat item (SN-trials, for signal-plus-noise). The test design then is as follows: 16 FTIs x 4 FTI categories x 2 Superposition levels x 2 View Difficulty levels x 2 Bag Complexity levels x 2 Bag Size levels x 2 trial types (SN and N-trials) resulting in the total of 2048 items.

Procedure

The X-Ray Bag Size Test was embedded into existing X-Ray Tutor training systems, in use at the airport. During screeners' regular training sessions the X-Ray Bag Size Test items would appear on screen in place of the usual X-Ray Tutor training items. Once the test was completed X-Ray Tutor would resume its usual training functions. With over two thousand items to solve, completing the test was likely to take several hours time. Screeners were not required to solve the test in one continuous session. Thus the vast majority solved the test in multiple sessions, spread out over a few days to several weeks. Apart from this the testing procedure in the X-Ray Bag Size Test is analogous to the procedure in the X-Ray ORT (see Experiment 1).

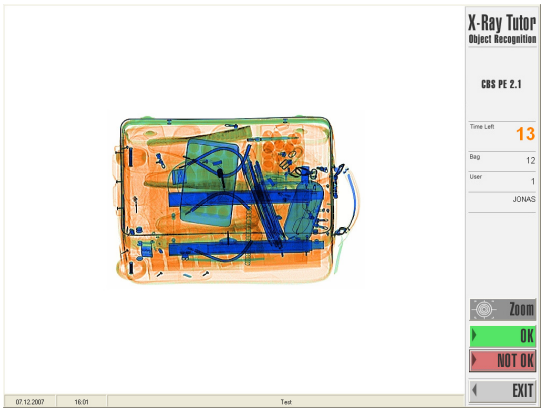

Figure 4.7: Screenshot of an X-Ray Tutor computer based test trial.

Statistical Analyses

To assess the individual image based factors' perceptual relevance for detection performance, as in Experiment 3, Pearson's product-moment correlation was used. To assess the perceptual relevance of all image based factors combined, multiple linear regression models were computed, as in Experiment 4.

4.7.2 Image Based Factors

For the X-Ray Bag Size Test some of the formulas for calculating image measurements had to be adapted.

FTI View Difficulty

The same formula was used as in Experiment 3.

Superposition

The same formula was used as in Experiment 3. However scatter plot inspection revealed a non-linear relationship between these Superposition values and d'. Thus an appropriate data transformation for Superposition had to be found. This transformation proved to be the negative logarithm of the negative Superposition. Superposition values had been defined such that they were always negative. Thus we had to negate these Superposition values before performing a logarithmic transformation. After the transformation we negated the values back again, thereby ensuring that the resulting variable reflect the proper ordinal relationships of the original Superposition values. This non-linear relationship between Superposition and detection performance which our transformation function revealed, indeed seems plausible: Intuitively we would expect additional Superposition in images with

already strongly superimposed FTIs to have a stronger effect on detection performance than in images where the FTI is hardly superimposed at all.

Clutter

Concerning the use of the Clutter formula within this experiment, please note that the summation term was standardized for Bag Size since Bag Size was implicitly included within the Clutter formula used in Experiment 3. Please compare the Bag Size formula below to the Bag Size formula under Equation 4.3 as applied in Experiment 3.

$$CL = \frac{\sum_{x,y} I_{hp}(x,y)}{BS} \tag{4.5}$$

For the details on the high pass filtering refer to Appendix A or Bolfing and Schwaninger (2007).

Opacity

The denominator of the Opacity formula was subject to a minor change: The threshold of the denominator (BS) changed from 252 to 254 since we no longer computed our formulas on jpg images. jpg images needed a lower threshold to prevent distorted values due to compression artifacts.

$$OP = \frac{\sum_{x,y}(I_N(x,y) < 64)}{BS} \tag{4.6}$$

Bag Size

Bag size is defined as the amount of pixels with a brightness value smaller than the brightness value 254, with 255 representing the maximum brightness.

$$\text{BS} = \sum_{x,y} (I_N(x, y) < 254) \tag{4.7}$$

4.7.3 Results

Pearson's Product-Moment Correlations

Table 4.6 shows an overview of the correlations between the individual image based factors and the measured performance d' for the four categories of threat items, guns, knives, IEDs and other. As in Experiment 3, View Difficulty is the strongest predictor for detection performance d' - with IEDs as the only exception: Detection performance for IEDs seems to be virtually independent of view. Superposition has strong correlations in all four categories of FTIs. Opacity sports comparatively low correlations and plays a more important role with knives and IEDs than with guns or other. Finally, correlations for Clutter and Bag Size are so low, that they do not reach statistical significance - the only exception being Bag Size in the case of knives.

Figure 4.8 shows the scatter plots of the four multiple linear regression models, with regression standardized predicted values on the x-axis plotted against measured detection performance d' on the y-axis. Table 4.7 gives an overview of the most important statistical values of the four models. As in study 4, all statistical models are highly significant in the overall goodness-of-fit verification statistics, indicated by the $R^2(adj)$ values. For guns and knives $R^2(adj)$ values are very high. They are yet higher than the predictions of the models based on human ratings or the models based on computed image factors in Experiment

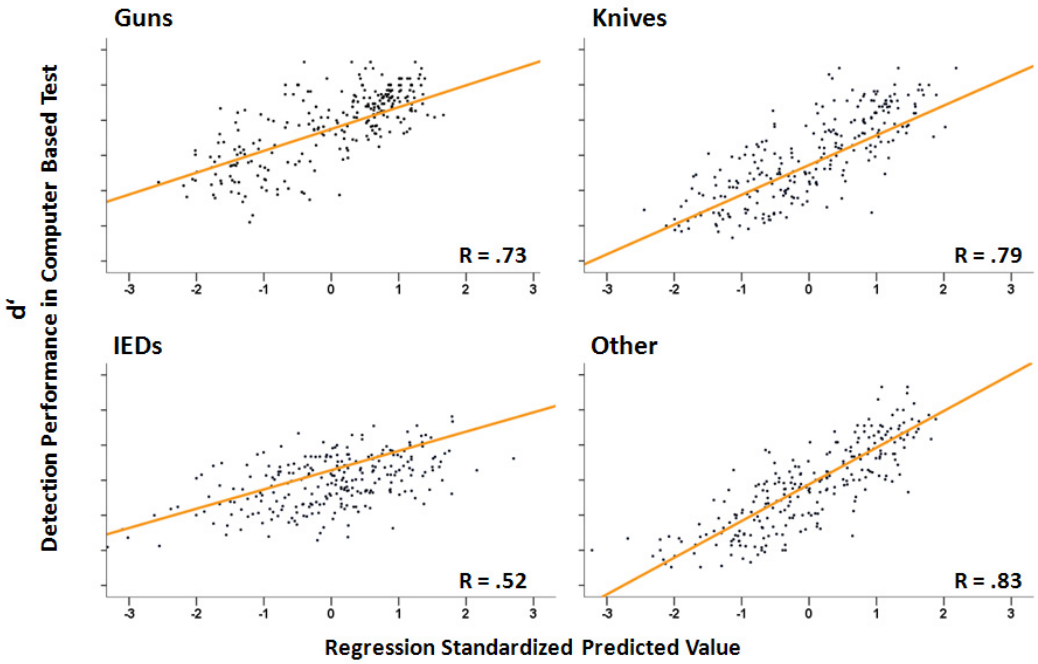

Figure 4.8: The four scatter plots from the models predicting d' on the basis of all disposable image based factors as predictors. Guns, knives, IEDs and other are displayed separately.

4. The model for the FTI Category other has the highest predictive value of all, with an $R^2(adj)$ of .680. For IEDs, with $R^2(adj)$=.266 the goodness of fit of the model is the lowest by far. To a considerable extent this will be due to the virtual non-correlation between detection performance for IEDs and FTI View Difficulty. All predictors are significant, mostly highly significant, except for two exceptions: Bag size which is never significant, and FTI View Difficulty which is not significant in the model for IEDs.

4.7.4 Discussion

The data strongly suggest that designing a new and larger test paid off. The already very high goodness-of-fit figures of the multiple linear regression models of Experiment 4 were surpassed by a considerable margin by the models of Experiment 5. Although intuitively

plausible, Bag Size turned out to play a negligible role in determining detection performance. No beta weights and only one correlation reaching significance lead to this conclusion. With regards to the importance of the different image based factors, for guns, knives and other a similar pattern emerged: FTI View Difficulty played the most important role, followed by Superposition. Opacity was a distant third. Clutter and Bag Size were largely negligible. For IEDs however there is one big difference: View difficulty does not reach significance in neither analysis.

4.8 General Discussion

Summarizing this article, we can state that our statistical model predicting detection performance d' as developed in Experiments 1-4 was very successfully approved by applying it on a new data set in Experiment 5. The results of Experiment 5 reveal a better predictive power of the model and more stable patterns in beta-weights than in Experiment 4. We assume that the reason for these differences lies in the larger data set used in Experiment 5. A closer look at the X-Ray ORT data reveals that there was a considerable ceiling effect, decreasing the overall variance in the data, which makes it very difficult to get stable data. Therefore we assume that the X-Ray ORT is too easy for expert screeners. Additionally it could be stated that the number of 19 very well trained screeners solving 64 SN-N pairs per FTI Category only in the X-Ray ORT (Experiment 1) is too small to get reliable statistical model prediction as opposed to the 63 screeners (trained and untrained) solving 256 SN-N pairs per FTI Category in the X-Ray Bag Size Test.

The formulae of image based factors introduced in this article promise to be powerful tools for both real world applications and further scientific research. The ability to make precise, perceptually plausible and valid measurements of image based factors leads to a plethora of opportunities. Some possible applications such as individually adaptive training algorithms in computer based training, or image difficulty estimation in threat image projection

(TIP) were already mentioned in the introduction. Furthermore our research should prove helpful for cost-benefit analyses in improving transport security. On the other hand highly predictive models as these are valuable for researching underlying mechanisms in human perception. For example, the virtual independence of Bag Size and detection performance indicates that visual search only plays a marginal role in the mental processing involved in this kind of detection task.

Table 4.6: Correlations between the Image Measurements and the measured Detection Performance d'

Guns: View Difficulty:	$r(256) = -.67$	$p < .001$
Superposition:	$r(256) = -.60$	$p < .001$
Opacity:	$r(256) = -.14$	$p < .05$
Clutter:	$r(256) = -.06$	$p = .351$
Bag Size:	$r(256) = -.03$	$p = .654$
Knives:		
View Difficulty:	$r(256) = -.69$	$p < .001$
Superposition:	$r(256) = -.49$	$p < .001$
Opacity:	$r(256) = -.28$	$p < .001$
Clutter:	$r(256) = -.10$	$p = .103$
Bag Size:	$r(256) = -.17$	$p < .01$
IEDs:		
View Difficulty:	$r(256) = -.09$	$p = .154$
Superposition:	$r(256) = -.40$	$p < .001$
Opacity:	$r(256) = -.29$	$p < .001$
Clutter:	$r(256) = -.03$	$p = .63$
Bag Size:	$r(256) = -.05$	$p = .473$
Other:		
View Difficulty:	$r(256) = -.79$	$p < .001$
Superposition:	$r(256) = -.68$	$p < .001$
Opacity:	$r(256) = -.16$	$p < .05$
Clutter:	$r(256) = -.05$	$p = .438$
Bag Size:	$r(256) = -.03$	$p = .600$

Model Summaries (Per Category)

	Predictors	Guns		Knives		IEDs		Other	
		β	p	β	p	β	p	β	p
Image Measurements	FTI View Difficulty	$-.52$	.000	$-.61$	.000	$-.09$	.112	$-.65$	.000
	-log(-Superposition)	$-.26$	.000	$-.21$	.000	$-.33$	.000	$-.21$	.061
	Opacity	$-.24$	.000	$-.36$	.000	$-.48$	.000	$-.30$	.000
	Clutter	.15	.025	.14	.029	.32	.000	.19	.003
	Bag Size	.02	.664	$-.06$	.183	.08	.184	.06	.158
	R^2	.53		.62		.27		.69	
	$adjusted R^2$	.52		.61		.25		.68	
	$F(5, 250)$	57		82		18		109	
	p	.000		.000		.000		.000	

Table 4.7: Statistical analysis tables of the models with the most important statistical values of the multiple linear regression analyses. Each of the four tables shows the statistical values of the verification of each regression coefficient separately in the rows. Additionally the model's overall goodness-of-fit verification values are given at the bottom row of each model. In both statistical models, the dependent variable is the d' obtained in Experiment 5.

Chapter 5

Appliance of the Model in the Political Decision-Making Process

5.1 The Impact of Image Based Factors and Training on Threat Detection Performance in X-ray Screening

In this study, two experiments are reported which investigated the relative importance of five different image based factors and one human factor (Training) in mediating threat detection performance of human operators in airport security X-ray screening. Experiment 1 was based on a random sample of roughly 16'000 records of threat image projection (TIP) data. TIP is a software function available on state-of-the-art X-ray screening equipment that allows the projection of fictional threat images (FTIs) into X-ray images of passenger bags during the routine baggage screening operation. Analysis of main effects showed that image based factors can substantially affect screener detection performance in terms of the hit rate (identification of FTIs). There were strong effects of FTI View Difficulty (rotation of FTIs) and Superposition of FTIs by other objects in the X-ray image of a passenger bag.

The amount of Opacity in the X-ray image of a passenger bag had a small although significant effect on detection performance. The two image based factors Clutter and Bag Size did not have a significant effect.

Experiment 2 was conducted using an offline-test in order to provide controlled and more detailed data for analyzing the image based factors from Experiment 1, as well as the human factor of training. In particular the individual factors' main effects on detection performance, main effects of all factors taken together and factor interactions were analyzed. In the test design the following image-based factors were varied systematically: Threat (FTI) Category (guns, knives, improvised explosive devices, other threats), View Difficulty, Superposition, Bag Complexity (a combination of Opacity and Clutter) and Bag Size. Data were collected from 200 screening officers at five sites across Europe. For screener training all five sites use the same computer-based training system. Consistent with the results obtained in Experiment 1, there were large main effects of Threat (FTI) Category, View Difficulty, and Superposition. Again consistent with Experiment 1, effects of Bag Complexity (Opacity and Clutter) and Bag Size were much smaller. In addition to Experiment 1, the number of computer based training (CBT) hours was available for each security officer participating in the study. Training turned out to be a key driver to improving threat detection performance in X-ray screening and seemed to mediate the effects of some image based factors.

This study was funded by the UK Department for Transport (DfT), on behalf of the ECAC[1] Technical Task Force. The study was conducted in collaboration with QinetiQ Ltd. and served as the scientific underpinning in the EU political decision-making process regarding a possible bag size restriction with in the European Union. Recommendations regarding the enhancement of human-machine interaction in X-ray screening are discussed.

[1] European Civil Aviation Conference

5.2 Introduction

Screening passenger bags for threat items using state-of-the art X-ray machines is an essential component of airport security. Previous works (Schwaninger, 2003b; Schwaninger, Hardmeier, & Hofer, 2005; Schwaninger, Michel, & Bolfing, 2007) have identified image based factors that affect human performance in X-ray screening tasks: object View Difficulty, Superposition by other objects and Bag Complexity (Opacity and Clutter). Recently the question has been raised whether Bag Size could be another image based factor that affects detection of threat items when visually inspecting X-ray images of passenger bags. In this study we determined effects and interactions of image based factors and human factors (amount of recurrent computer-based training). In addition, with empirically based conclusions regarding the importance of the Bag Size variable, by itself as well as in relation with other performance relevant factors, this study provided the scientific basis for a political decision-making process regarding the improvement of aviation security.

Two experiments are reported. Experiment 1 is based on threat image projection (TIP) data. Experiment 2 is based on an off-line computer based test, which allows investigating the combined effects of image-based factors, effects of training as well as factor interactions. The use of these two methods to answer the same research question will ensure that the overall approach is complementary. Both methods have their own strengths and weaknesses: TIP data give high ecological validity but low experimental control; off-line computer based tests using controlled stimuli allow more experimental control, but less ecological validity. If both methods provide the same answer to the research question, this can be taken as stronger evidence that the findings are genuine, and not simply an artefact of the particular method used.

The two experiments both follow the paradigm using computer algorithms to estimate image based factors that influence threat detection performance in X-ray screening. This paradigm was developed at University of Zurich and presented at ICRAT 2006 in Belgrade (Bolfing, Michel, & Schwaninger, 2006a) and published (Schwaninger, Michel, &

Bolfing, 2007; Bolfing, Michel, & Schwaninger, 2006b; Schwaninger, Michel, & Bolfing, 2005). None of these papers used TIP data for analysis, which ensures high ecological validity. Experiment 2 is based on a much larger data set than the previous studies, augmenting reliability. The inclusion of Bag Size and training as additional factors is completely novel within this paradigm. Since threat detection performance in aviation security X-ray screening depends on the X-ray images but also on the human screeners - the final decision makers - human factors should not be neglected in a comprehensive model whose goal is to explain the X-ray threat detection process.

5.2.1 Image Based Factors

Schwaninger (2003b) and Schwaninger, Hardmeier, and Hofer (2005) have identified three image based factors which affect threat detection by X-ray screeners: View Difficulty, Superposition, and Bag Complexity (see Figure 5.1).

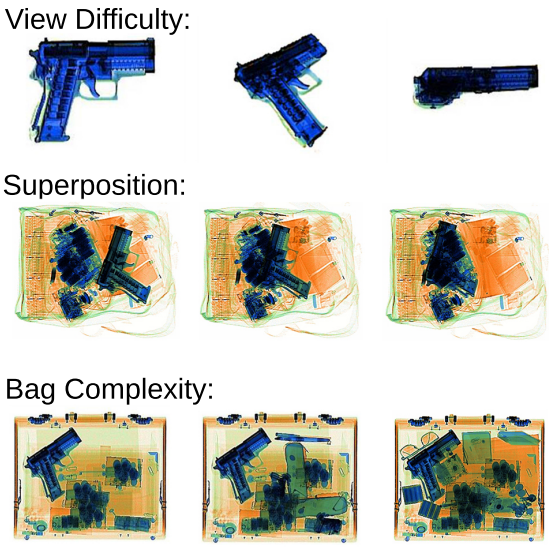

Figure 5.1: Illustration of the three basic image based factors suggested by Schwaninger (2003b) and Schwaninger, Hardmeier, and Hofer (2005)

The concepts of these image based factors have been mathematically modeled (Schwaninger,

Michel, & Bolfing, 2007). See Bolfing and Schwaninger (2007) for the latest version. View difficulty is modeled as a statistically calculable value between 0 and 1 named FTI View Difficulty. Superposition and Bag Complexity are modeled as image processing measurements with Bag Complexity being split up into Clutter and Opacity. The introduction of the image based factor Bag Size in this study necessitated normalization of earlier implementations of Clutter and Opacity regarding Bag Size. Formulae and short descriptions of the underlying concepts are specified in Bolfing and Schwaninger (2007) in Appendix A.

5.3 Threat Image Projection (TIP) χ^2 Analysis: Experiment 1

5.3.1 Method

Threat Image Projection (TIP) Data

In order to ensure high ecological validity, we decided to analyze data from threat image projection (TIP) systems. TIP is a software function of state-of-the-art X-ray screening equipment used at security checkpoints in airports, nuclear power plants, navigation docks etc. In aviation security TIP distinguishes between cabin baggage screening (CBS) and hold baggage screening (HBS). In CBS, guns, knives, improvised explosive devices (IEDs) and other threats are subject to identification and confiscation. In HBS, the focus rests mainly on IEDs and dangerous goods such as gasoline containers or diver lamps. The current investigation is confined to CBS. In CBS TIP, fictional threat items (FTIs) are occasionally projected into X-ray images of passenger bags during the routine baggage screening operation. A sufficiently large sample of TIP events allows statistically reliable measurements of detection performance of human operators (X-ray screeners) on-the-job (Hofer & Schwaninger, 2005) and thus with high ecological validity.

The data basis of this study consists of a random sample of 16'329 TIP events that have been routinely recorded on-the-job with approximately 700 professional X-ray screeners throughout the first half of 2007 at a large European airport. We decided to apply χ^2 analyses to each image based factor separately to measure its impact on detection performance in terms of hit rate (i.e. correctly judging a bag as being NOT OK).

χ^2 Analysis

To compare the effects on detection performance of the independent variables[2] FTI View Difficulty, Superposition, Opacity, Clutter and Bag Size, the following procedures were applied to the TIP data described above. A histogram was created for each independent variable (image based factor). For each variable the upper and lower 2.5% of the cases in the data were excluded to remove outlier data from the analysis. Furthermore this made possible the definition of five equidistant bins with at least 100 data points each (TIP events).
Hit rates were calculated for each of the five equidistant bins to run χ^2 tests with the null hypothesis H0 that the hit rates are equal across bins. Effect size analysis based on Cohen (1988) was used to compare the effect sizes of the different independent variables. For detailed information on χ^2 statistics see for example Coolican (2004).

5.3.2 Results

The results below are listed separately for each image based factor introduced above (see Bolfing and Schwaninger (2007) for further information and formulae). Each of the following subsections begins with a graphical illustration of the image based factors' effects on the threat detection performance measure hit rate. The x-axes show the five equidistant

[2]The variables correspond to the continuously represented variables used in the multiple regression analysis in Experiment 2 (see Figure 5.8)

bins into which the whole data range was subdivided. Low values are on the left, high values on the right. The y-axes show the hit rates of the image based factors' bins. For reasons of confidentiality hit rates cannot be given explicitly, but the hit rate scales are reasonably chosen and kept constant throughout the whole document.

Following the graphical illustrations (Figures 5.2-5.6), statistical test values are given in Tables 5.1-5.5. χ^2 statistics can be interpreted as follows: the larger the $\chi^2(df, N)$ value the larger the effect. Additionally χ^2 effect sizes w are given. Again, the larger the effect size, the larger the effect. However, please be aware that χ^2 and w values do not state the direction of the effect.

To summarize the χ^2 analysis results a bar plot graphic is provided at the end of this section illustrating the χ^2 effect sizes of the five image based factors on the hit rate (see Figure 5.7). The image based factors are arranged such that their effects decrease in size.

Figure 5.2 illustrates the large impact of FTI View Difficulty on human detection performance in terms of hit rate. This is partly due to the fact that objects are more difficult when depicted from an unusual viewpoint (see Figure 5.1). Other factors contributing to this large impact are the Threat Category of the object and the training of human operators (see Experiment 2). Figure 5.3 illustrates the large effect of Superposition on detection performance. Figure 5.4 shows the significant but relatively small influence of Opacity on detection performance in terms of hit rate. Here the question arises whether it is Opacity as a perceptual concept that does not have much influence on threat detection performance, or whether the image measurement formula of Opacity is not properly modeled. Figure 5.5 illustrates the hit rates of the five Clutter bins. There is no significant effect of Clutter on detection performance. As with Opacity, the question arises whether it is the concept of Clutter that does not influence hit rates in TIP, or whether the computational model of Clutter needs to be improved. Figure 5.6 shows the effect of Bag Size on hit rate in TIP. As with Clutter, the effect of Bag Size on detection performance does not reach statistical significance.

FTI View Difficulty

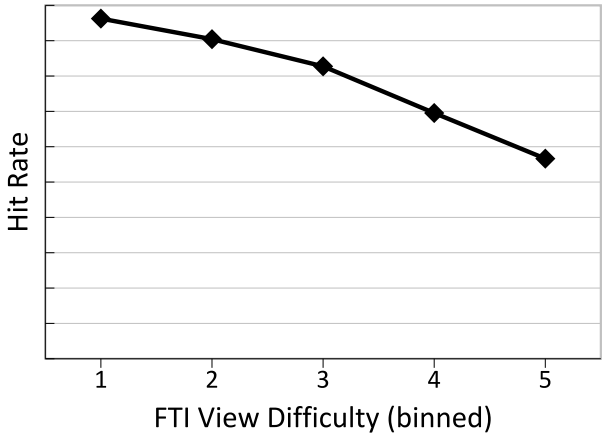

Figure 5.2: Illustration of the impact of FTI View Difficulty on hit rate.

Table 5.1: χ^2 Analysis Results: FTI View Difficulty

χ^2 value	$\chi^2(4, N = 13'541) = 198.04$
Significance	Highly significant: $p < .001$
χ^2 effect size	$w = .12$

Superposition

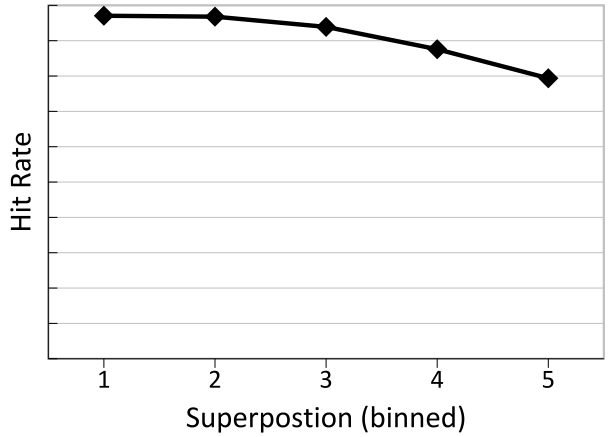

Figure 5.3: Illustration of the impact of Superposition on hit rate.

Table 5.2: χ^2 Analysis Results: Superposition

χ^2 value	$\chi^2(4, N = 13'713) = 72.98$
Significance	Highly significant: $p < .001$
χ^2 effect size	$w = .07$

Opacity

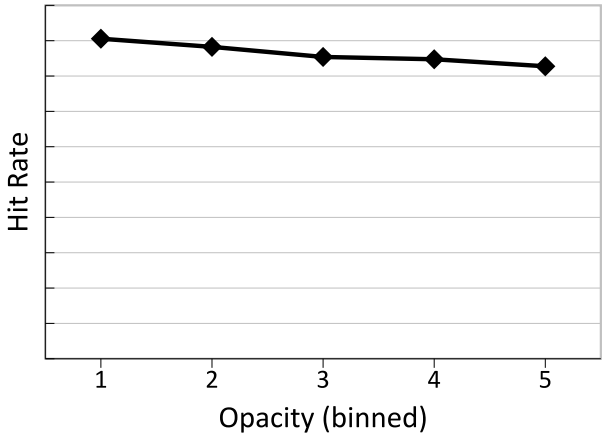

Figure 5.4: Illustration of the impact of Opacity on hit rate.

Table 5.3: χ^2 Analysis Results: Opacity

χ^2 value	$\chi^2(4, N = 13'718) = 9.90$
Significance	Significant: $p < .05$
χ^2 effect size	$w = .03$

Clutter

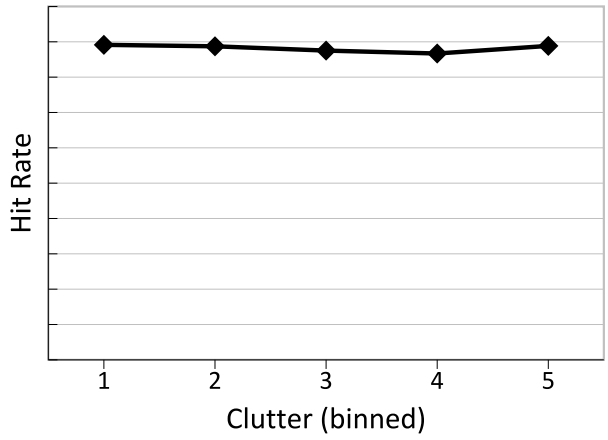

Figure 5.5: Illustration of the impact of Clutter on hit rate.

Table 5.4: χ^2 Analysis Results: Clutter

χ^2 value	$\chi^2(4, N = 13'726) = 0.98$
Significance	Not significant: $p = .913$
χ^2 effect size	$w = .01$

Bag Size

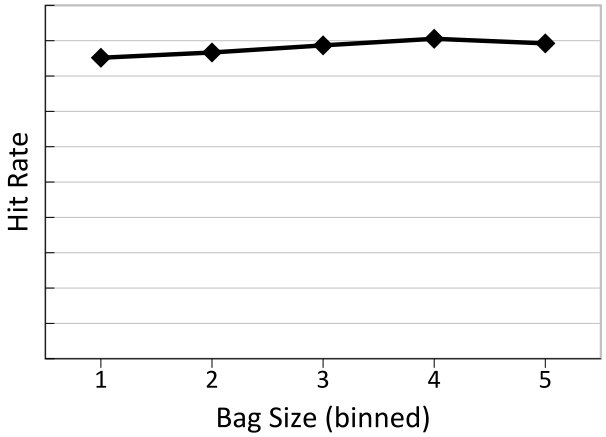

Figure 5.6: Illustration of the impact of Bag Size on hit rate.

Table 5.5: χ^2 Analysis Results: Bag Size

χ^2 value	$\chi^2(4, N = 13'758) = 4.45$
Significance	Not significant: $p = .348$
χ^2 effect size	$w = .02$

Comparison of the χ^2 Effect Sizes

In Figure 5.7, the effect sizes w are compared. The factor FTI View Difficulty has the highest effect size with $w = .12$, while Clutter shows the lowest effect size with $w = .01$. The factors Opacity, Bag Size and Clutter show small effect sizes. The effects of Clutter and Bag Size did not reach statistical significance.

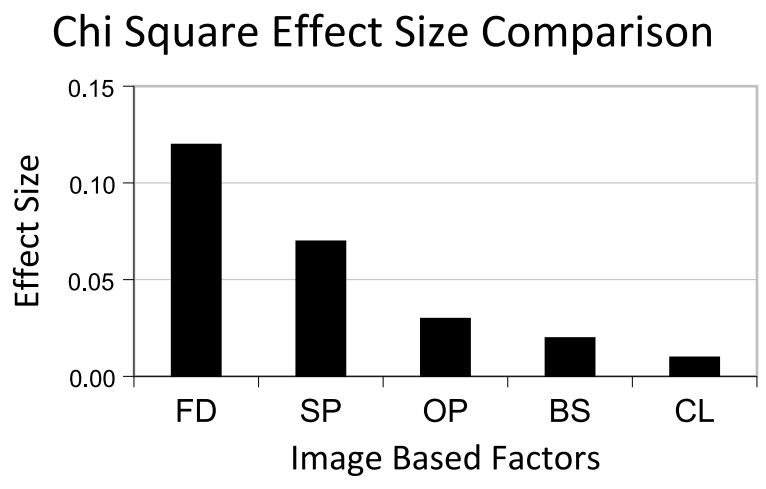

Figure 5.7: Comparison of the effect sizes among the image based factor.

5.3.3 Discussion

The results obtained in Experiment 1 are consistent with earlier findings. Schwaninger, Hardmeier, and Hofer (2005) found that View Difficulty, Superposition and Bag Complexity affect screener performance. Schwaninger, Michel, and Bolfing (2007) replicated these results (see Chapter 2). Using similar image measurements as in Experiment 1, they measured similar effects for FTI View Difficulty, Superposition, Opacity and Clutter. However, several caveats are necessary to qualify the appropriateness of the results obtained in Experiment 1. Firstly, an analysis of auto-archived bags indicated that, as would be anticipated, it is likely that TIP aborts are selectively eliminating certain bags (e.g. small bags rather

than large bags) from the TIP image set, and thus reducing their presence. Secondly, it is not always clear how closely aligned TIP scores are with the specific operational situations encountered when threats are deliberately hidden in difficult bags. But most importantly, in Experiment 1 only main effects were analyzed. In order to gain a more complete picture it is important to conduct a more controlled experiment in which main effects in combination and their interactions can be measured reliably. This was done in Experiment 2.

5.4 Off-line Computer Based Test: Experiment 2

5.4.1 Method

Participants

200 X-ray screeners from five European airports with varying amounts of training in X-ray image interpretation agreed to participate in this study.

Stimuli

The stimuli consisted of 1024 X-ray images of passenger bags containing a threat item (SN trials; signal-plus-noise) and the same 1024 bags not containing any prohibited items (N trials; noise). The SN trial images were created by projecting fictional threat items (FTIs) into 1024 X-ray images of bags. FTIs for the study were eight visually similar pairs of each of four types of threat items: guns, knives, improvised explosive devices (IEDs), and 'other' threats. Images of cabin baggage were captured from X-ray machines at a European airport using the auto-archive function. The images were revised by three airport security supervisors to remove inappropriate images (e.g. images containing more than one bag,

images containing incomplete bags, bags containing prohibited items or liquids, etcetera). This procedure resulted in 7606 bag images. Additional review by the QinetiQ team (our collaboration team) resulted in a total of 6659 bag images from which the 1024 bags needed for the study were drawn. The final 1024 bags used for the study were chosen through a process of projecting the relevant FTIs into the bags such that the variables of interest would be orthogonal in the stimulus set. Several full sets of 2048 images (the 1024 images containing the FTIs, and the same images without FTIs) were created. The one with the most desirable properties in terms of variable orthogonality was chosen for use in the study.

Design

The study employed a 4 (FTI Category: guns, knives, IEDs, other) x2 (View Difficulty: easy, difficult) x2 (Superposition: low, high) x2 (Bag Complexity: low, high) x2 (Bag Size: small, large) x2 (image type: SN, N) within-participants design. Since there were 16 FTIs in each Category, this design results in a total of 16x4x2x2x2x2x2 = 2048 images which were to be presented to the screeners. The images were presented to the screeners in a random order in multiple testing sessions of 20 minutes each. As dependent variable the detection performance measure d' (Green & Swets, 1966) was used. This measure provides a more valid estimate of detection performance than the hit rate alone because it takes the hit rate and the false alarm rate into account (see (Hofer & Schwaninger, 2004) for different measures of X-ray detection performance). Since the off-line test showed each bag once with a threat item and once without, accurate measurements of hit and false alarm rates could be obtained.

5.4.2 Results

Data were analyzed in two ways. Firstly, by treating the variables FTI View Difficulty, Superposition, Opacity, Clutter, and Bag Size as continuous, a linear regression was employed to assess the main effects of each image based factor on threat detection performance separately. A multiple linear regression was used to examine the main effects together. Additionally, we calculated a linear regression with hours of recurrent computer based training prior to testing as predictor. In order to examine main effects as well as interactions between the variables, the variables FTI Category, View Difficulty, Superposition, Bag Complexity and Bag Size, all transformed into discrete variables with two parameter values low and high, were used in an analysis of covariance (ANCOVA). Training hours served as the covariate variable in the ANCOVA. Figure 5.8 shows the way in which the continuous and discrete variables are related to each other. Due to a high inter-correlation and a test design that demands independence of its variables, Opacity and Clutter were encoded into the single discrete variable Bag Complexity. FTI Category and View Difficulty were encoded into a single continuous variable because it is not sensible to encode either variable directly into a continuous variable. Instead we defined the variable FTI View Difficulty as the difficulty - as measured in threat detection performance (d') - screening officers had in solving a specific threat item in a specific view (easy or difficult) across all other conditions (i.e. Superposition, Bag Complexity and Bag Size).

Linear Regression and Multiple Linear Regression

The regression analyses will help us understand the direct relationship between image based factors and d', as well as training hours and d'. Figure 5.9 shows the relative effect sizes, the absolute values of the correlations with the dependent variable d', for the individual variables. For Superposition and training hours a logarithmic transformation was applied. This transformation was necessary in order to achieve a linear relationship

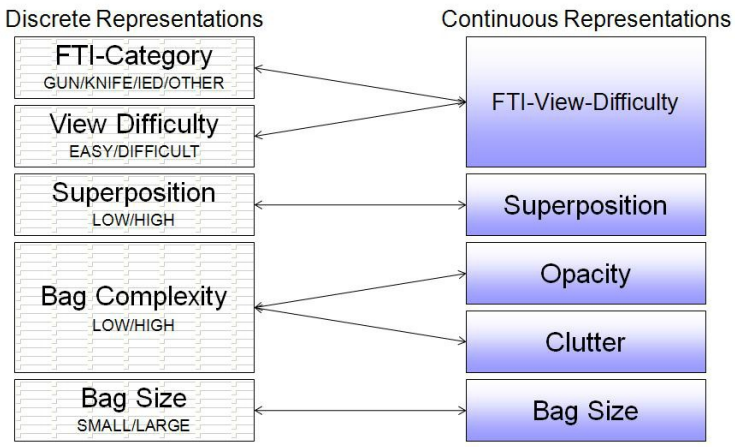

Figure 5.8: Illustration of the relationship between discrete and continuous representations of the independent variables

between Superposition and detection performance d'. With .70, .63 and .58, FTI View Difficulty, training hours and Superposition all have very high effect sizes. Opacity has a moderate to small effect size with .22, Clutter and Bag Size have very small effect sizes with .05 and .07, respectively. Except for Clutter, all correlations are statistically significant.

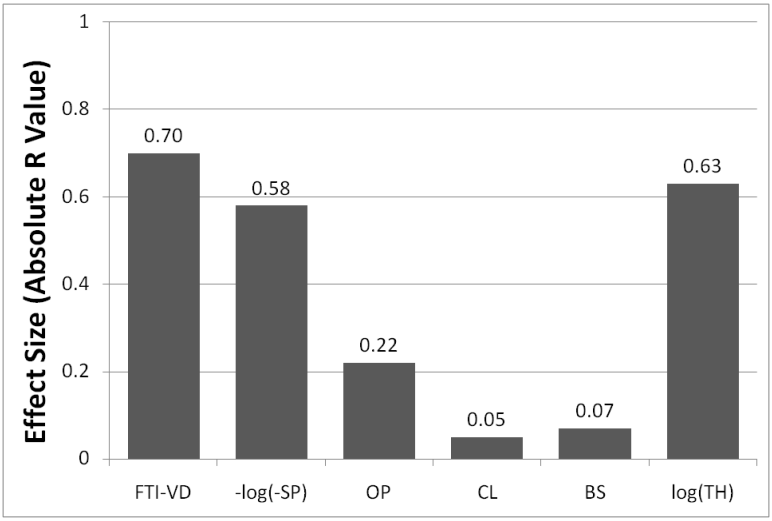

Figure 5.9: Illustration of effect sizes R

Figure 5.6 shows the results of the multiple linear regression with all image based factors: FTI View Difficulty, Superposition (logarithmically transformed), Opacity, Clutter and Bag Size. It shows the overall effect size, again the absolute value of the correlation R, of all the image based factors taken together. With $R = 0.77$ the effect size is very high. The effect size of the only human factor analyzed (hours of recurrent computer based training), with $R = 0.63$, is also large. We can see that in the multiple linear regression model the factor Bag Size is the only one not reaching statistical significance. Put another way: In the presence of the other image based factors Bag Size did not lead to a statistically significant change in detection performance in our experiment. As shown in Figure 5.10 adding Bag Size to the linear model only leads to a minimal increase of its effect size from $R = 0.772$ to $R = 0.773$.

Model Summaries (All Categories)

	Predictors	Beta weights β	Significance p
Image Based Factors	FTI View Difficulty	.568	.000
	logSuperposition	-.227	.000
	Opacity	-.366	.000
	Clutter	.223	.000
	Bag Size	.030	.193
	$R^2 = .60$, adjusted $R^2 = .60$, $p < .000$		

Table 5.6: Tabular summary of the general multiple linear regression models for all threat item categories. Standardized beta weights and p-values.

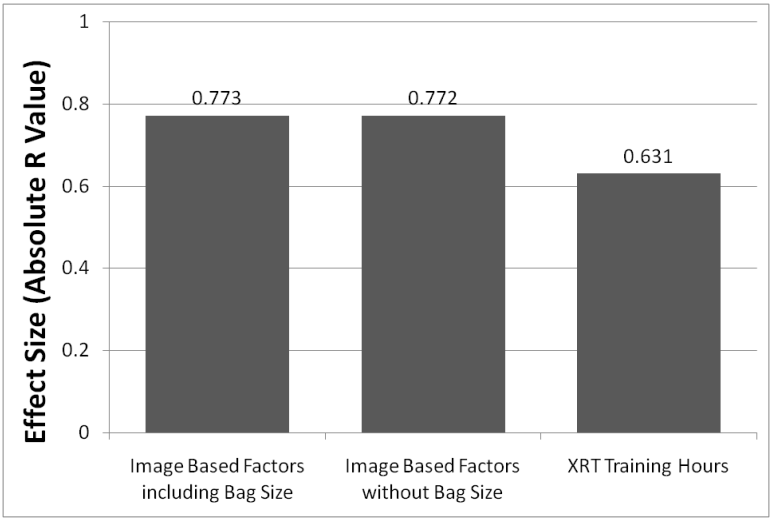

Figure 5.10: Combined effect size of image based factors and effect size of training

ANCOVA

A repeated measures analysis of covariance (ANCOVA) was conducted to analyze the main effects of image based factors, their interactions and their interactions with training. As can be seen in the main effects summary of Figure 5.11 the repeated measures ANCOVA leads to only a slightly different pattern with regards to effect sizes than the linear regression analyses. These differences are due to the fact that, in contrast to the linear regression models, in the ANCOVA analysis effects of the covariate training hours are isolated from the effects of image based factors. Furthermore, in the ANCOVA inter-individual differences between screening officers ('screener variance') are taken into account. Superposition shows the largest effect size (η^2), followed by FTI Category, Bag Complexity and View Difficulty. The main effect of Bag Size is clearly smaller than the main effect of any other image based factor. Training hours has noteworthy interactions with FTI Category and View Difficulty. These interactions make sense, since we know from other studies that training can lead to comparatively larger performance increases for items that are comparatively difficult for novices (Koller, Hardmeier, Hofer, & Schwaninger, 2008) - for example improvised explosive devices (threat item category) or difficult views (View Difficulty). There is also

a small interaction of training with Bag Size, indicating that well trained screening officers are less affected by effects of Bag Size. Figure 5.12 gives an overview of the 10 largest interactions in the ANCOVA. All in all over 30 interactions reached statistical significance. Since the effect sizes of most interactions are very small we decided only to report interactions $\eta^2 \geq .07$. The interaction of View Difficulty with Threat Category can at least partly be explained by the fact that detection performance of improvised explosive devices - unlike guns or knives - is largely independent of viewpoint. The interaction of Superposition with View Difficulty indicates that with difficult viewpoints Superposition plays a larger role in determining detection performance than with easy views. The interaction of Superposition with Threat Category indicates that some threat item categories are more sensitive to Superposition than others. For example, from the regression analysis above we know that Superposition effects are higher with knives than with guns.

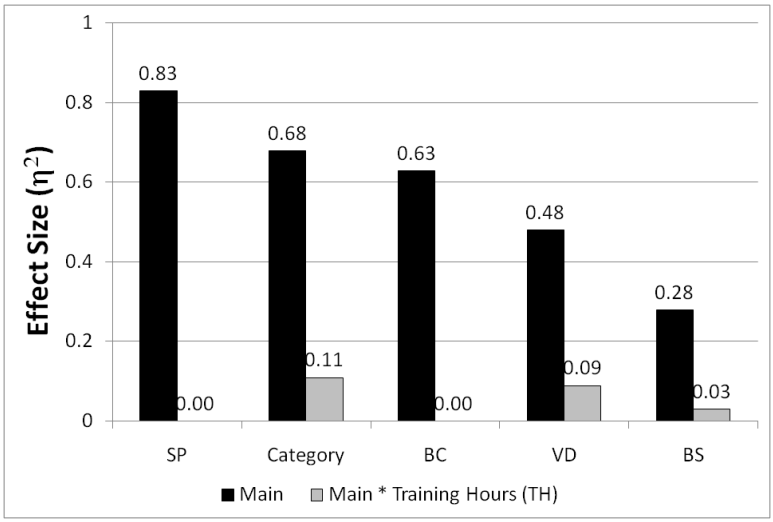

Figure 5.11: Illustration of ANCOVA main effects and interactions with the covariate training hours

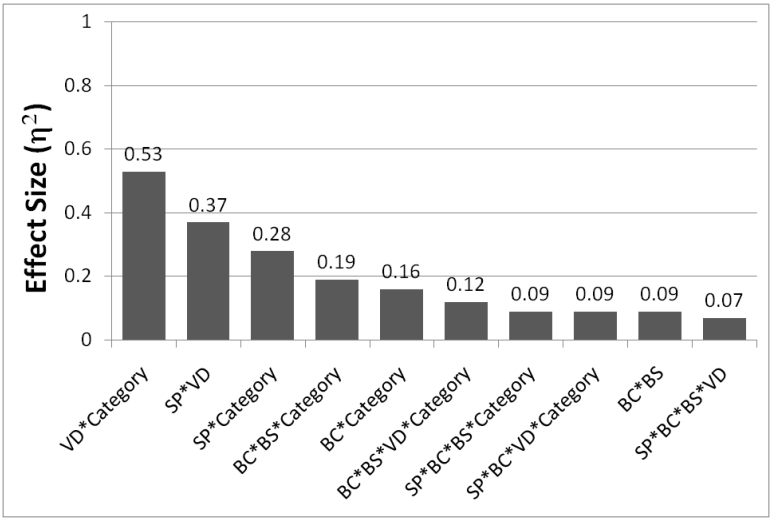

Figure 5.12: Illustration of the the ten largest ANCOVA interactions

5.4.3 Discussion

With an overall correlation of .77 the linear modeling of detection performance with image based factors has a very high explanatory power. Superposition, although not always with the largest effect size, has shown the most robust effects on detection performance. Interestingly and in contrast to what one might have expected based on the results of the regression analyses, the variable Bag Complexity (a combination of Opacity and Clutter) showed a large effect size in the ANCOVA. Apart from this, the ANCOVA results reflect the regression analysis results closely, both in main effects and interactions. Threat Category and View Difficulty had considerable interactions with the covariate training hours. This shows that training is particularly effective in the case of difficult item categories such as IEDs and for difficult viewpoints. Bag size, although intuitively plausible as relevant factor, turned out to play only a minor role in determining threat detection performance. The same is true for Clutter.

5.5 General Discussion

There were large main effects of View Difficulty and of FTI Category in all of the analyses, as expected. The same was true for Superposition and Bag Complexity (to a bigger extent for Opacity than for Clutter). Clearly, these factors need to be taken into account in any future work on performance-relevant image based factors. When looking at the influence on detection performance of all image based factors together, there is no statistically significant effect of Bag Size. When using a more sophisticated model of data analysis including main effects of FTI View Difficulty, Superposition, Bag Complexity, Bag Size and the interactions of these variables, there is a small effect of Bag Size. In Experiment 2 we were also able to examine the effect of the number of CBT training hours on threat detection performance. The key finding from the study is that the effect size for this variable was large, and seemed to mediate the effect of some image based factors on threat detection. Clearly, training is a key driver to improving threat detection performance in X-ray screening, and more work needs to be done to establish exactly which image based factors screeners need to be trained in to give the best improvements in threat detection accuracy.

5.6 Recommendations for Improving Human-Machine Interaction in X-Ray Screening

5.6.1 FTI View Difficulty and Superposition

The factor FTI View Difficulty refers to the fact that the identification of threat objects, as objects in general, is highly dependent on their viewpoint as well as on properties of the very object itself. Current X-ray screening equipment provides only one X-ray image per passenger bag. More recent technology can provide multiple views of a bag. Figure 5.13

illustrates how such new systems might be able to reduce the detection problems due to View Difficulty and Superposition. Objects that are superimposed by other objects from one perspective may be clearly visible from another one. Furthermore, training is an important tool in lessening detrimental effects on detection performance of difficult views. Our ANCOVA analysis has supported earlier findings that training leads to particularly large improvements in detection performance for difficult views (Koller et al., 2008).

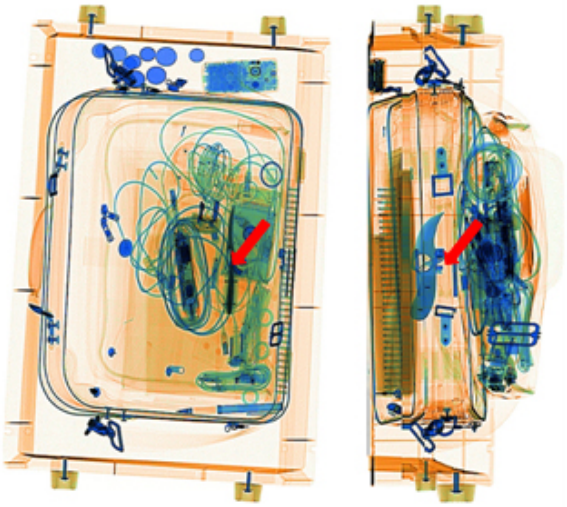

Figure 5.13: Illustrative example of how multi-view systems can help improving detection performance in spite of undesirable View Difficulty and Superposition effects.

5.6.2 Opacity

The image based factor Opacity refers to the amount of opaque areas in an X-ray image. X-ray systems with higher penetration have the potential to reduce detection problems due to Opacity. In addition, it is possible to implement image measurement algorithms in X-ray equipment that warn the human operator (X-ray screener) with a ”dark alarm”, which would be triggered by opaque areas that are deemed too large or dense for unassisted human interpretation. Manual search would follow when a dark alarm was indicated.

5.6.3 Screener Selection and Training

A very important approach to face the problem of improving threat detection performance in X-ray screening consists in screener selection and screener training. The psychological literature provides evidence that figure ground segregation (related to Superposition) as well as mental rotation (related to View Difficulty) are visual abilities that are fairly stable within a person. For example Hofer, Hardmeier, and Schwaninger (2006) and Hardmeier et al. (2006b) have shown that using computer based object recognition tests in a pre-employment assessment procedure can help to increase detection performance of screeners substantially.

In addition to stable abilities, there are several aspects of visual knowledge relevant to X-ray image interpretation. Knowledge based factors such as knowing which objects are dangerous or prohibited and what they look like in X-ray images are trainable. Training also has beneficial effects on screeners' abilities to deal with certain image based factors. For example, training particularly improves the ability to deal with difficult views. Computer-based training can be a powerful tool to improve X-ray image interpretation competency of screeners Koller et al. (2008); Schwaninger, Hofer, and Wetter (2007); Ghylin et al. (2006).

References

Annual review of civil aviation 2005. (2006). In *International civil aviation organization* (Vol. 61, p. 9).

Bolfing, A., Michel, S., & Schwaninger, A. (2006a). Assessing image difficulty in x-ray screening using image processing algorithms. In *Proceedings of the 2nd international conference on research in air transportation, icrat 2006, belgrade, serbia and montenegro, june 24-28* (pp. 253–258).

Bolfing, A., Michel, S., & Schwaninger, A. (2006b). A statistical approach for automated image difficulty estimation in x-ray screening using image processing algorithms. In *Proceedings of the 4th international aviation security technology symposium, washington, d.c., usa, november 27 - december 1* (pp. 384–388).

Bolfing, A., & Schwaninger, A. (2007). Measurement formulae for image-based factors in x-ray imagery. In *Vicoreg technical report, november 26.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Earlbaum Associates.

Coolican, H. (2004). *Research methods and statistics in psychology* (4th ed.). London: Hooder & Stoughton.

Gale, A., Mugglestone, M., Purdy, K., & McClumpha. (2000). Is airport baggage inspection just another medical image? In *Medical imaging: Image perception and performance. progress in biomedical optics and imaging* (Vol. 1(26), pp. 184–192).

Ghylin, K. M., Drury, C. G., & Schwaninger, A. (2006). Two-component model of security inspection: application and findings. In *16th world congress of ergonomics, iea 2006, maastricht, the netherlands, july* (pp. 10–14).

Green, D. M., & Swets, J. A. (1966).
In *Signal detection theory and psychophysics* (pp. 187–194). New York: Wiley.

Hardmeier, D., Hofer, F., & Schwaninger, A. (2005). The object recognition test ort - a reliable tool for measuring visual abilities needed in x-ray screening. In *Ieee iccst proceedings* (Vol. 39, pp. 189–192).

Hardmeier, D., Hofer, F., & Schwaninger, A. (2006b). Increased detection performance in airport security screening using the x-ray ort as pre-employment assessment tool. In *Proceedings of the 2nd international conference on research in air transportation, icrat 2006, belgrade, serbia and montenegro, june 24-28* (pp. 393–397).

Hofer, F., Hardmeier, D., & Schwaninger, A. (2006). Increasing airport security using the x-ray ort as effective pre-employment assessment tool. In *Proceedings of the 4th international aviation security technology symposium, washington, d.c., usa, november 27 - december 1* (pp. 303–308).

Hofer, F., & Schwaninger, A. (2004). Using threat image projection data for assessing individual screener performance. In *Ieee iccst proceedings* (pp. 303–308).

Hofer, F., & Schwaninger, A. (2005). Reliable and valid measures of threat detection performance in x-ray screening. In *Wit transactions on the built environment* (Vol. 82, pp. 417–426).

Koller, S., Hardmeier, D., Hofer, F., & Schwaninger, A. (2008). Investigating training, transfer, and viewpoint effects resulting from recurrent cbt of x-ray image interpretation. In *Journal of transportation security.*

Koller, S., & Schwaninger, A. (2006). Assessing x-ray image interpretation competency of airport security screeners. In *Proceedings of the 2nd international conference on research in air transportation, icrat 2006, belgrade, serbia and montenegro, june 24-28, 2006.*

Krupinski, E. A., Berger, W. G., Dallas, W. J., & Roehrig, H. (2003). Searching for nodules: What features attract attention and influence detection? In *Academic radiology* (Vol. 10(8), pp. 861–868).

Liu, X., Gale, A., Purdy, K., & Song, T. (2006). Is that a gun? the influence and features of bags and threat items on detection performace. In *Contemporary ergonomics, proceedings of the ergonomic society* (pp. 17–22).

Madden, D. J., Gottlob, L. R., & Allen, P. A. (1999). Adult age differences in visual search accuracy: Attentional guidance and target detectability. In *Psychology and aging* (pp. 683–694).

Mahfouz, M. R., Hoff, W. A., Komistek, R. D., & Dennis, D. A. (2005). Effect of segmentation errors on 3d-to-2d registration of implant models in x-ray images. In *Journal of biomechanics* (Vol. 38(2), pp. 229–239).

McCarley, J. S., Kramer, A. F., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual skills in airport-security screening. In *Psychological science* (Vol. 15, pp. 302–306).

Riegelnig, J., & Schwaninger, A. (2006). The influence of age and gender on detection performance and the criterion in x-ray screening. In *Proceedings of the 2nd international conference on research in air transportation, icrat 2006, belgrade, serbia and montenegro, june 24-28, 2006* (pp. 403–407).

Schwaninger, A. (2003b). Evaluation and selection of airport security screeners. In *Airport* (Vol. 2, pp. 14–15).

Schwaninger, A. (2004b). Computer based training: a powerful tool to the enhancement of human factors. In *Aviation security international, feb/2004* (pp. 31–36).

Schwaninger, A. (2005b). Increasing efficiency in airport security screening. In *Wit transactions on the built environment* (Vol. 82, pp. 407–416).

Schwaninger, A. (2006b). Airport security human factors: From the weakest to the strongest link in airport security screening. In *Proceedings of the 4th international*

aviation security technology symposium, washington, d.c., usa, november 27 - december 1 (pp. 265–270).

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. In *Ieee iccst proceedings* (Vol. 38, pp. 258–264).

Schwaninger, A., Hardmeier, D., & Hofer, F. (2005). Aviation security screeners visual abilities & visual knowledge measurement. In *Ieee aerospace and electronic systems* (Vol. 20(6), pp. 29–35).

Schwaninger, A., Hofer, F., & Wetter, O. E. (2007). Adaptive computer-based training increases on the job performance of x-ray screeners. In *Proceedings of the 41st carnahan conference on security technology, ottawa, october 8-11, 2007.*

Schwaninger, A., Michel, S., & Bolfing, A. (2005). Towards a model for estimating image difficulty in x-ray screening. In *Ieee iccst proceedings* (Vol. 39, pp. 185–188).

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements. In *Proceedings of the 4th symposium on applied perception in graphics and visualization, acm press, new york, usa* (pp. 123–130).

Sluser, M., & Paranjape, R. (1999). Model-based probabilistic relaxation segmentation applied to threat detection in airport x-ray imagery. In *Proceedings of the 1999 ieee canadian conference on electrical and computer engineering.*

Ying, Z., Naidu, R., & Crawford, C. R. (2006). Dual energy computed tomography for explosive detection. In *Journal of x-ray science and technology* (Vol. 14(4), pp. 235–256).

Appendix A

Formulary

A.1 Image Measurement Formulary

A.1.1 Method

All image measurements developed for this purpose are based on theoretical considerations. Different algorithm parameters were optimized by maximizing the correlations between the image-based factors estimates and detection performance measures derived from earlier X-Ray ORT findings from x-ray screening experts.

FTI View Difficulty

Even with the aid of 3D volumetric models, it is not (yet) possible to satisfyingly determine the degree of a 3-dimensional rotation (view difficulty) of a physical threat item automatically from its 2-dimensional x-ray image (Mahfouz et al., 2005). Additional difficulties regarding image segmentation arise from the very heterogeneous backgrounds of x-ray images, compare (Sluser & Paranjape, 1999). Therefore, this image based factor is

not(yet) being calculated by image processing, but statistically from X-Ray ORT detection performance data obtained in Experiment 1. Equation A.1 shows the general a posteriori statistical formula for FTI View Difficulty as applied in chapters 1-4. Equation A.3 is the more specific formula used in TIP data analysis presented in chapter 5.

$$\text{FtiVD}_{OVj} = \frac{\sum_{i=1,j\neq i}^{N_{OV}} (max(\text{DetPerf}) - \text{DetPerf}_{OVi})}{N_{OV} - 1} \tag{A.1}$$

$$\text{FtiVD}_{OVj} = \frac{\sum_{i=1,j\neq i}^{4} (4.65 - d'_{OVi})}{3} \tag{A.2}$$

$$\text{FtiVD}_{OV} = \frac{\sum_{i=1}^{N_{OV}} \text{MissRate}_{OVi}}{N_{OV}} \tag{A.3}$$

The general FTI View Difficulty formula depicted in Equation A.1 reads as follows: FtiVD_{OVj} represents the indexed abbreviation of the FTI View Difficulty of the X-ray image (or SN-N X-ray image pair, in question, indexed by j. Basically, the FTI View Difficulty formula is just the average of all detection performance values containing a fictional threat item in a certain view. The index OVj represents a certain FTI object (index O) in a certain view (index V) presented N_{OV} times in a test or in TIP. In order to adjust the resulting value to the direction representing difficulty the measured detection performance ($DetPerf_{OVi}$) is subtracted from the theoretical maximum detection performance value. In case of the analyzed threat detection performance measure being d' $max(DetPerf)$ is set to 4.65. In case of the hit rate or A' being the analyzed detection performance $max(DetPerf)$ is set to 1.0. In short, FTI View Difficulty is calculated by averaging the inverted threat detection performance across all X-ray images containing a certain FTI, but excluding the X-ray image in question from averaging. In case of the analyzed test being X-Ray ORT, for example,

the number of images included in the averaging is three (Equation A.2). N_{OV} is four (2 bag complexities x 2 superpositions) whereby one image is excluded ($N_{OV} - 1$). The exclusion of the one item in questions necessitates from avoiding circular arguments in multiple linear regression analyses. Equation A.3 represents the FTI View Difficulty formula as it was used in TIP data analysis in chapter 5. Due to the enormous size of the TIP data set for FTI View Difficulty estimation (191677 TIP events), we refrained from excluding the one image in question from averaging. In our TIP data analysis each of the 2010 FTIs was projected 95 times during the half-year period on average. In this case, the effect of a possible circular argument diminished to a fraction of about 1%. Further, the reason why the miss rate was used instead the generally used expression $max(DetPerf) - DetPerf_{OV}$ is that in TIP, no false alarm rates can be recorded. Therefore, the detection performance used is the hit rate only. Correspondent to the general formula, $1 - hitrate = missrate$, the theoretical maximum detection performance minus the actually measured detection performance.

It is important to understand that this concept of FTI View Difficulty is not just reflecting the degree of rotation of an object. In that case there would be two parameter values for all threat exemplars only. View Difficulty as it is conceptualized here reflects innate view difficulty attributes unique to each exemplar view separately.

Superposition

This image based factor refers to how much the pixel intensities at the location of the FTI in the threat bag image differ from the pixel intensities at the same location in the same bag without the FTI. Equations A.4 - A.5 all show slightly different implementations of the image based factors measurements of Superposition. All formulas are based on the same principle. In all equations $I_{SN}(x, y)$ denotes the pixel intensity values of a threat item image and $I_N(x, y)$ denotes the pixel intensity values of the corresponding bag image not containing any threat item. Equation A.4 shows the Superposition formula as implemented

in the study in chapter 2. Equation A.5 shows the Superposition formula as implemented in chapters 3 to 5. The only difference as opposed to Equation A.4 is the constant term C which is arbitrary. Basically, the subtraction of Superposition as in Equation A.5 was introduced to invert the direction of the Superposition value such that it fits with detection performance difficulty of the images. The constant term C only serves as an aesthetic correction to make values positive. Therefore C can be freely chosen depending on the Superposition value range from Equation A.4. For pure calculation purposes we suggest to set C to 0. In chapters 3 and 4, where enough data points were available to estimate the true relationship between Superposition and detection performance, a log-transform was applied to the Superposition values as in Equations A.4 - A.5. The corresponding formulae can easily be reconstructed and are not given here.

$$\text{SP} = \sqrt{\sum_{x,y} \left(I_{SN}(x,y) - I_N(x,y)\right)^2} \tag{A.4}$$

$$\text{SP} = C - \sqrt{\sum_{x,y} \left(I_{SN}(x,y) - I_N(x,y)\right)^2} \tag{A.5}$$

It should be noted that this mathematical definition of superposition is dependent on the size of the threat item in the bag. For further development of the computational model it is conceivable to split up superposition and the size of the threat item into two separate image based factors. Measurement of superposition would require having both the bag with the FTI and without. For both applications mentioned in the introduction, this is possible with current TIP and computer-based training (CBT) technology. In TIP, the FTI, its location, the bag with and without the FTI are recorded. In several CBT systems, the same information is recorded and stored, too.

To date, Superposition values are applied to greyscale images only. But since colour coding is an important improvement in state-of-the-art X-ray machines, the Superposition measurement algorithms will be applied to the single colour coding channels in the near future.

Together with the implementation of our image based factor Colour Saliency, which is not subject to this dissertation, we expect this to further enhance our image based factors models.

Clutter

This image based factor is designed to express bag item properties like its textural unsteadiness, disarrangement, chaos or just clutter. In terms of the bag images presented, this factor is closely related to the amount of items in the bag as well as to their structures in terms of complexity and fineness. The method used in this study is based on the assumption, that such texture unsteadiness can be described mathematically in terms of the amount of high frequency regions.

$$\mathrm{CL} = \frac{\sum_{x,y} I_{hp}(x,y)}{BS} \tag{A.6}$$

$$\begin{aligned} \text{where} \quad I_{hp}(x,y) &= I_N * \mathcal{F}^{-1}(hp(f_x, f_y)) \\ &= \mathcal{F}^{-1}(\mathcal{F}(I_N \cdot hp(f_x, f_y)) \end{aligned}$$

Equation A.6 shows the image measurement formula for Clutter. It represents a convolution of the empty bag image (N for noise) with the convolution kernel derived from a high-pass filter in the Fourier space. I_N denotes the pixel intensities of the harmless bag image. $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. $hp(f_x, f_y)$ represents a high-pass filter in the Fourier space.

Clutter formula high-pass filter where f_x and f_y are its frequency components, f is its cut-off frequency and where d is its fall-off.

$$hp(f_x, f_y) = 1 - \frac{1}{1 + \left(\frac{\sqrt{f_x^2 + f_y^2}}{f}\right)^d} \tag{A.7}$$

This high-pass filter represents a 2-D matrix in the Fourier frequency space. Therefore an inverse Fourier transformation is applied to transform it into a convolution kernel in the spatial domain.

Unfortunately, to date all experiments and applications including Clutter as a predictor variable regarding threat detection performance showed quite disappointing effects of Clutter. We are actually quite confident that it is not our theoretical concept of Clutter but its computational implementation which must be subject to a major revision. This is quite a challenge. What seems to be a highly complex image from a computer perspective might be easily perceived by humans and vice versa. As an example, images of human faces may consist of highly complex pixel patterns, but are perceived by humans very fast and reliably. In informatics, on the other hand, a huge amount of compression algorithms exist that allow to highly reduce complexity in structures that humans perceive as highly complex. Tracking down this gap between computational complexity and complexity as perceived by human minds is far from being examined and understood in scientific literature.

Opacity

The image based factor Opacity, designated Transparency in earlier studies, reflects the extent to which X-rays are able to penetrate objects in a bag. This depends on the specific material density of these objects. These attributes are represented in X-ray images as different degrees of luminosity. Heavy metallic materials, such as lead for example, are known to be very hard to be penetrated by X-rays and therefore appear as dark areas on X-ray images.

$$\text{OP} = \frac{\sum_{x,y} \big(I_N(x,y) < 64\big)}{\text{BS}} \tag{A.8}$$

Equation A.8 shows the image measurement formula for Opacity. $I_N(x,y)$ denotes the pixel intensities of the harmless bag. threshold is the pixel intensity threshold beneath which the pixels are counted. The implementation of the image measurement for the image based factor Opacity is simply achieved by counting the number of pixels being darker than a certain threshold relative to the bag's overall size. In this dissertation the threshold was consistently set to 64, which equals one fourth of the whole pixel intensity range. The denominator in Equation A.5 equals Bag Size (BS) as described in the next section.

Bag Size

As already mentioned under Opacity, the formula for Bag Size just expresses the fact that all non-white pixels of the X-ray image are counted. In this dissertation we set the threshold for non-white pixels to 254 instead of 255, due to possible image capture artifacts of the X-ray machines.

$$\text{BS} = \sum_{x,y} (I_N(x,y) < 254) \tag{A.9}$$

Appendix B

Addendum - Curriculum Vitæ

Personal information

Surname, First name	**Bolfing, Anton**
Address	Müllerstrasse 43, 8004-Zürich, Switzerland
E-mails	a.bolfing@psychologie.uzh.ch
Nationality	Swiss
Date of birth	12.07.1975, Schwyz, Switzerland
Gender	male

Work experience

Dates	since January 2007
Occupation or position held	Doctoral student
Main activities and responsibilities	statistical modelling, image processing, psychophysics
Name and address of employer	University of Zurich, Department of Psychology, General Psychology (Cognition), Visual Cognition Research Group, Binzmühlestrasse 14/22, CH-8050 Zürich
Type of business or sector	Aviation security
Dates	since January 2007
Occupation or position held	Doctoral student
Main activities and responsibilities	EU project management, data analysis, reporting, presenting
Name and address of employer	Max Planck Institute for Biological Cybernetics, Speemannstr. 38, D-72076 Tübingen
Type of business or sector	Aviation security; cognitive and computational psychophysics department
Dates	2004 - 2006
Occupation or position held	Research assistant
Main activities and responsibilities	Data analysis (statistical modeling), quality control management
Name and address of employer	University Zurich, Klosbachstrasse 107, CH-8032 Zürich
Type of business or sector	Aviation security
Dates	2003 - 2004
Occupation or position held	Research assistant
Main activities and responsibilities	Data analysis (statistical modeling), quality control management
Name and address of employer	Applied Psychological Science Solutions GmbH
Type of business or sector	Aviation security; cognitive psychology

Dates	January - August 1999
Occupation or position held	Internship
Main activities and responsibilities	IT migration
Name and address of employer	Zürcher Kantonalbank, Neue Hard 11, 8005 Zürich
Type of business or sector	Logistics

Education and training

Dates	1999 - 2006
Title of qualification awarded	A Statistical Approach for Image Difficulty Estimation in X-Ray Screening Using Image Processing Algorithms
Principal subjects/ occupational skills covered	Studies in Psychology, Neurophysiology and Philosophy
Name and type of organization providing education	University Zurich
Level in national or international classification	MSc UZH, Psychologist[1]
Dates	Fall 2004
Principal subjects/ occupational skills covered	Assessment Center, competitor analysis, office
Name and type of organization providing training	xcg executive consulting group, executive assessment center
Level in national or international classification	Internship
Dates	Fall 2003
Principal subjects/ occupational skills covered	Psychophysics, saccade adaptation
Name and type of organization providing education	Ludwig Maximilian University, Munich
Level in national or international classification	Internship
Dates	1996 - 1998
Principal subjects/ occupational skills covered	Studies in Physics and Mathematics
Name and type of organization providing education	ETH Zürich
Level in national or international classification	not completed

Personal skills and competences

Mother tongue	**German**
Other languages	English (good), French (basic), Italian (basic)
Social skills and competences	Team Work: Always very good experiences and feedback in a lot of temporal occupations
Organisational skills and competences	During the work at VICOREG acquiring good skills in coordinating projects, dealing with project partners and customers
Technical skills and competences	Competent in quantitative and qualitative research methods, experimental designing, data analysis
Computer skills and competences	Fully competent with most Microsoft computer programs (MS Windows, MS Excel, MS Word, MS Powerpoint, MS Project, MS Access), Matlab programming (numerical computation), LaTeX (typesetting), SPSS & R (statistics), Adobe Photoshop, and some Java programming
Artistic skills and competences	Guitar, some painting, design
Driving licence	Swiss car driving license

Additional Information

Teaching Experience

Laboratory Practical Course Lecture at the Department of Industrial and Organisational Psychology, University of Zurich (Experimental-psychologisches Praktikum der Abteilung Arbeits- & Organisationspsychologie der Universitt Zürich)

Collaborations

Max Planck Institute for Biological Cybernetics, Tübingen, Germany (Prof. H.H. Bülthoff);
Zurich State Police, Airport Division (F. Hofer);
Airport Security Clearance, Bulgaria (A. Yankov)

Publications

Peer Reviewed Articles

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., & Hay L. (2008). The impact of image based factors and training on threat detection performance in X-ray screening. Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008, Fairfax, Virginia, USA, June 1-4, 2008, 317-324

Schwaninger, A., Michel, S., & Bolfing, A. (2007). A statistical approach for image difficulty estimation in x-ray screening using image measurements. Proceedings of the 4th Symposium on Applied Perception in Graphics and Visualization, ACM Press, New York, USA, 123-130

Bolfing, A., Michel, S., & Schwaninger, A. (2006b). A statistical approach for automated image difficulty estimation in x-ray screening using image processing algorithms. Proceedings of the 4th International Aviation Security Technology Symposium, Washington, D.C., USA, November 27 December 1, 2006, 384-388

Bolfing, A., Michel, S., & Schwaninger, A. (2006a). Assessing image difficulty in x-ray screening using image processing algorithms. Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006, 253-258

Schwaninger, A., Michel, S., & Bolfing A. (2005). Towards a model for estimating image difficulty in x-ray screening. IEEE ICCST Proceedings, 39, 185-188

Technical Reports

Bolfing, A., & Schwaninger, A. (2007). Measurement formulae for image-based factors in x-ray imagery. VICOREG Technical Report, November 26, 2007

Conference Abstracts

Michel, S., Bolfing, A., & Schwaninger, A. (2005). Modelling image based factors in aviation security x-ray screening. Poster presented at the 47. Tagung experimentell arbeitender Psychologen, April 4-6, Regensburg, Germany

Michel, S., Bolfing, A., Hofer, F., & Schwaninger, A. (2004). Towards a Psychophysically Plausible Model of X-Ray Threat Detection Performance. Poster presented at the 2nd LIDOKO (Lizentianden und Doktorandenkongress), Zurich

Invited Lectures and Presentations

Schwaninger, A., Bolfing, A., Halbherr, T., Helman, S., Belyavin, A., & Hay L. (2008). The impact of image based factors and training on threat detection performance in X-ray screening. Proceedings of the 3rd International Conference on Research in Air Transportation, ICRAT 2008, Fairfax, Virginia, USA, June 1-4, 2008, 317-324

Bolfing, A. (2008). The Impact of Image Based Factors and Training on Threat Detection Performance in X-ray Screening - a study involving 5 European airports. InterTAG, May 27-29, Toulouse

Bolfing, A. (2007). Image-based factors that affect the detection of threat items in x-ray images defined, VIA Meeting, September 26-28, Berlin

Bolfing, A. (2007). VIA project screener's survey results (workpackage 4), VIA Meeting, September 26-28, Berlin

Bolfing, A. (2007). VIA project screener's survey results (workpackage 5), VIA Meeting, September 26-28, Berlin

Bolfing, A., Michel, S., & Schwaninger, A. (2006). Assessing image difficulty in x-ray screening using image processing algorithms. Proceedings of the 2nd International Conference on Research in Air Transportation, ICRAT 2006, Belgrade, Serbia and Montenegro, June 24-28, 2006, 253-258